Identification of key biomarkers in neonatal sepsis by integrated bioinformatics analysis and clinical validation

Rui Yan, Tao Zhou*

Department of Pediatrics, Shenzhen Baoan Women's and Children's Hospital, Jinan University, Shenzhen 518102, China

ARTICLE INFO

Keywords:
- Neonatal sepsis
- Diagnostic biomarker
- Immune cell infiltration
- Inflammatory response
- GSVA
- LASSO model

ABSTRACT

Background: Neonatal sepsis (NS) is a systemic inflammatory response to severe pathogenic infections, and is a major cause of high morbidity and mortality in newborns. Currently, there is a lack of efficient diagnostic technology to accurately and rapidly diagnose NS, and the precise pathogenesis of NS has yet to be fully elucidated. The present study aimed to identify the optimal biomarkers in the progression of NS.

Methods: The differentially expressed genes (DEGs) between NS and controls in the discovery datasets were screened. Gene set variation analysis (GSVA) was used to enrich the changes in biological functions and pathways in sepsis patients compared to healthy individuals. The differences in immune cell infiltration between these two groups were assessed using CIBERSORT. Furthermore, LASSO algorithm and ROC analysis were performed to identify and evaluate the gene signature.

Results: A total of 85 upregulated and 40 downregulated overlapping DEGs were screened in sepsis samples. The GSVA results indicated that DEGs largely contributed to upregulated inflammation and metabolism-related processes, and suppressed adaptive immune responses in NS. Markedly lower infiltration of most types of immune cell was observed in sepsis patients, except for some innate immune cells. Moreover, 57 genes with AUC >0.9 in both discovery sets were selected and applied to a LASSO model. Using this model, a seven-gene signature was acquired, which was validated in the discovery and independent validation sets. Five genes among the gene signatures with optimal diagnostic performance were obtained and further validated in clinical samples using RT-qPCR. Finally, three genes SLC2A3, OSCAR, and CD3G were identified as key biomarkers for NS.

Conclusions: Our findings will provide novel insights into the pathogenesis of NS, and the potential biomarkers may have promising application values for its early detection and therapeutic intervention.

1. Introduction

Neonatal sepsis (NS) refers a life-threatening and dysregulated inflammatory response when newborns are infected by pathogenic bacteria (Weiss et al., 2020). Pathogenic bacteria, including bacteria, viruses, and fungi, enter the blood circulation system of newborns, grow and multiply in the circulation, and generate toxins that spread throughout the body, eventually resulting in severe systemic inflammatory response (Shane et al., 2017). NS is a critical illness in the neonatal period and is associated with a high morbidity and mortality worldwide, seriously threatening the health of infants (Kissoon and Uyeki, 2016). Based on data from the last decade, it is estimated that globally neonatal sepsis was 2202 for every 10000 livebirths, with a mortality of 11–19%, meaning that 3 million newborns suffer from sepsis every year (Fleischmann-Struzek et al., 2018). The severity of NS varies with the economic and medical levels of different countries, which creates a huge economic and mental burden on society.

NS is a serious infectious disease caused by pathogens. Once the host response to infection is dysregulated, the inflammatory response will damage its own cells and tissues, eventually leading to multiple organ failure and even death (Ziesmann and Marshall, 2018). In the development of sepsis, the early excessive inflammatory response triggered by infection often evolves into subsequent immune dysfunction. However, the investigation of pathophysiological process of neonatal sepsis, as well as the therapeutic strategy, is still lacking. It is worth noting that more than half of newborns die within the first three days after the onset of sepsis. There is also evidence that mortality in the acute onset of NS is closely related to a dysregulation of the pro-inflammatory response (Khaertynov et al., 2017; Pietrasanta et al., 2019), which indicates that...
the acute pro-inflammatory phase is crucial in the early diagnosis and therapeutic intervention of NS. Since sepsis is not easily diagnosed based on clinical manifestations alone, laboratory diagnostic techniques are constantly being updated (Patel and McElvania, 2019). At present, the gold standard for the diagnosis of NS in most countries is blood culture, which takes a long time, with test results generally available after 48 h. Due to the small amount of blood collection, false negatives often occur, thus missing the optimal time window for clinical treatment (Tarai et al., 2012). Some blood biomarkers, such as C-reactive protein (CRP), Procalcitonin (PCT), and Interleukin (IL)-6, have been used to assist in the predictive and early diagnosis of neonatal sepsis (Altuhana et al., 2011; Sharma et al., 2018). These biomarkers have their own advantages and disadvantages, and their specificity and sensitivity still need to be further improved for clinical applications (Gillilain and Bhandari, 2017; Iroh Tam and Bendel, 2017). In a word, despite advances in the diagnosis of sepsis, there is currently no single test standard that can accurately and rapidly diagnose neonatal sepsis (Chauhan et al., 2017).

Therefore, it is particularly urgent to explore the molecular changes that occur during the pathogenesis of NS and discover effective and novel biomarkers for early diagnosis and therapeutic options. At present, bioinformatics analysis combined with gene expression profiling has become a powerful approach to reveal the pathophysiology of sepsis, and has been widely applied to the discovery of molecular targets for disease treatment (Wang and Liotta, 2011; Chen et al., 2019).

In the present study, we employed a comprehensive bioinformatics method to analyze genome-wide transcriptomic data in neonatal sepsis to identify emerging gene biomarkers and validate them in clinical samples. We hope that these key gene markers will supply novel insights into the underlying mechanism of neonatal sepsis, and will be used as early diagnostic biomarkers.

2. Materials and methods

2.1. Data sources and preprocessing

Three gene expression datasets (GSE25504, GSE26440 and GSE69686) were downloaded from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database, which include gene expression profiles and corresponding clinical information of neonatal sepsis samples. GSE25504 dataset contained 44 normal samples and 42 neonatal sepsis samples, which were based on the GPL6947 platform (Illumina Human HT-12 V3.0 array), GPL570 platform (HG-U133_Plus_2) Affymetrix Human Genome U133 Plus 2.0 Array and GPL13667 platform. GSE25504 was submitted by Dickinson et al., GSE26440, consisting of 10 normal samples and 16 patient samples, was based on the GPL570 platform (HG-U133_Plus_2) and submitted by Wong et al. The platform of GSE69686 was GPL20292 (Illumina Custom Affymetrix Human Transcriptome Array), which included 58 normal samples and 64 neonatal sepsis samples. The raw data were processed in R package, and the robust multichip average (RMA) algorithm was used for correction and normalization of the data. Sixty-three samples in GSE25504, which were processed on the GPL6947 platform, and 26 samples in GSE26440 were selected as the discovery set, while the remaining samples in GSE25504 (based on GPL570 platform and GPL13667 platform, named as GSE25504-V) and samples in GSE69686 were taken as the validation set. The discovery set was mainly used for model construction, discovery, and the identification of enriched pathways, whereas the validation set was mainly used to evaluate the model and validate the hub genes. In this analysis, infants aged of less than 3 months were defined as neonates. Drugs targeting gene markers are predicted using the Drug Gene Interaction Database (DGIdb).

2.2. Selection and analysis of differentially expressed genes

The differentially expressed genes (DEGs) between sepsis patients and healthy controls in GSE25504 and GSE26440 were performed using the limma package in R (Ritchie et al., 2015). Herein, the adjusted P-values (adj. P) and Benjamini-Hochberg correction method were applied in this analysis. Genes with |log (fold change)| ≥ 1 and adj. P-value < 0.05 were identified as DEGs. Then, an online Venn software (http://jvenn.toulo use.inra.fr/app/) was applied to generate the common DEGs, which were upregulated or downregulated in the same direction in the two datasets. In addition, unsupervised hierarchical clustering analysis of common DEGs between sepsis patients and controls was also conducted and visualized using the heatmap package (https://hiplot.com.cn/basic /heatmap).

2.3. Protein-protein interaction (PPI) network construction

To assess the interactions among common genes in NS, a PPI network was constructed using the online STRING database (https://string-db.org). A combined score >0.4 was defined as the cut-off criterion. Then, the PPI network topology of the overlapping DEGs was displayed using Cytoscape software (version 3.7.2) (Shannon et al., 2003). The CytoHubba plug-in was installed in Cytoscape and employed to calculate the degree of interaction between the nodes (genes) in the PPI network. The nodes were ranked according to their degrees of connectivity.

2.4. Gene set variation analysis

The gene set variation analysis (GSVA) package in R, which can be downloaded at http://www.bioconductor.org, was applied to display the enrichment of GO terms and KEGG pathways in neonatal sepsis patients (Hänzelmann et al., 2013). GSVA can be regarded as a pathway-level difference analysis. The GSVA calculation method is used to calculate the pathway enrichment score of each sample in the database and obtain a pathway enrichment data. Then the GSVA score was analyzed by limma package in R to determine the differences in pathway activation between the sepsis and control samples. Expression scores with |logFC| ≥ 0.5 and P < 0.05 were considered statistically significant.

2.5. Immune cell infiltration evaluation

CIBERSORT is an analytical tool for extracting features from RNA-Seq data and modifying the RNA expression data into the proportions of various immune cells using a deconvolution algorithm. Normalized gene expression data in GSE25504, GSE26440, and GSE69686 were used to discriminate 22 subtypes of human immune cells between sepsis patients and controls. The scores of infiltrating immune cells of these there datasets were calculated based on CIBERSORT, which derives a p-value to evaluate the results. P < 0.05, indicated that the obtained fractions of immune cell subsets were accurate. Pearson's correlation was used to assess the correlations between hub genes and the infiltration of different immune cell types. Not all 22 immune cell subsets can be obtained from gene data of newborn samples, therefore, 20 subtypes were included in this study.

2.6. Identification of gene signature by ROC analysis and LASSO model

The receiver operating characteristic (ROC) analysis and area under the curve (AUC) values were applied to evaluate the discriminatory power of each gene in neonatal sepsis versus controls using the pROC package in R. Common genes with AUC value >0.90 in both GSE25504 and GSE26440 were selected for further screening via the LASSO algorithm model. The Lasso algorithm performs via a penalty function, compressing the coefficient of variables and setting some regression coefficients to zero in order to achieve the purpose of variable selection, thus obtaining a more refined model. To further narrow the range of the candidate genes, the LASSO algorithm was applied to construct an optimal model with fewer genes utilizing the R software and the glmnet package (version 4.1). Ten-fold cross-validation with binomial deviance was used to select the penalty parameters (optimal lambda). Based on the
best lambda value, the candidate diagnostic gene for sepsis, named the gene signature, can be achieved by refitting the model with the gene expression data. Then, logistic regression combined with ROC analysis based on multiple genes was carried out to assess the effectiveness of the gene signature in differentiating sepsis from healthy controls (diagnostic performance) in the validation set. Finally, the specificity and sensitivity of the diagnostic value of each gene among the gene signatures were calculated by the ROC curve. Genes with AUC > 0.75 in both the discovery and validation sets were identified as hub genes.

2.7. Gene network construction and functional analysis of the co-expression genes

Pearson's correlation analysis was used to count the interactions of the hub genes with all the identified DEGs in the discovery set. If Pearson's r ≥ 0.70 and P < 0.05 were selected as the cutoff criteria. The co-expression network was established and visualized by Cytoscape software. Functional enrichment interpretation of co-expression genes was performed using Metascape. In this bioinformatics tool, for each given gene sequence, the following ontology sources including GO biological processes, KEGG pathway, and Reactome gene sets et al. were conducted for biological process and pathway enrichment analysis. Genome-wide genes were used as the enrichment background. Selected terms with P < 0.01, enrichment factor > 1.5, and a minimum count of 3 were grouped into clusters according to their degree of similarity. P-values were obtained via the cumulative hypergeometric distribution, and the most statistically significant term in a cluster was selected to represent the cluster. The subgroup of enriched terms was used to build the network plot, which was visualized using Cytoscape. In the network, different nodes represent enriched terms, and nodes with a similarity > 0.3 were connected by edges.

2.8. Clinical samples

Neonatal peripheral blood samples of 22 neonatal sepsis patients and 16 uninfected control neonates from Shenzhen Baoan Women’s and Children’s Hospital of Jinan University were collected for the verification of the candidate key genes. The attending clinician determines a neonatal for sepsis based on the expert consensus on the diagnosis and management of neonatal sepsis (version 2019). The diagnostic criteria are shown in Supplementary Table S3. The study was approved by the Medical Ethics Committee of Shenzhen Baoan Women’s and Children’s Hospital of Jinan University. All examinations were informed by the guardian of each newborn. Reverse transcription-quantitative PCR (RT-qPCR) was performed to validate the expression of key genes in the current study.

2.9. Gene expression analysis by RT-qPCR

Total RNA from whole blood samples was isolated using the EasyPure® Blood RNA Kit (Transgen Biotech, Beijing, China) according to the manufacturer’s protocols. Purity and concentration of RNA were measured by a NanoDrop one C UV-Vis spectrophotometer (Thermo Fisher Scientific, Madison, USA). Total RNA was reverse transcribed for the synthesis of first-strand cDNA using HiScript® III All-in-one RT SuperMix (Vazyme, Nanjing, China) following manufacturer’s instructions. Quantitative PCR was conducted using Taq Pro Universal SYBR qPCR Master Mix (Vazyme) and in an ABI 7500 real-time PCR machine (ABI Biosystems). The primer sequences were shown in Supplementary Table S1. GAPDH was used as the reference gene and gene expression was normalized to the expression of GAPDH, obtaining the ΔCT value. The mRNA levels relative to GAPDH were calculated by 2^ΔCT.

2.10. Construction of a diagnostic nomogram

To further predict the clinical diagnostic value of the screened biomarkers, we constructed a nomogram using the R package version 4.1.2. The discriminative performance of the nomogram was assessed with the concordance index (C-index) and the ROC curve. A C-index or the AUC value of ROC curve equal to 0.5 suggests that the nomogram has no predictive ability, while a C-index or AUC value of 1.0 indicates that the model has a great ability to distinguish patients from healthy controls. A greater the AUC value in this model reflects better diagnostic value.

2.11. Statistical analysis

The statistical significance of immune cell infiltration and gene expression analysis were evaluated using a two-tailed unpaired Student’s t-test. All statistical analyses, except specially noted, were performed using GraphPad Prism Software v.7.03 (GraphPad Software Inc., San Diego, CA, USA). Statistical significance was defined as P < 0.05.

3. Results

3.1. Identification of overlapping DEGs in neonatal sepsis

A detailed flowchart of this study is shown in Figure 1. To identify NS-associated genes, differentially expressed genes (DEGs) between NS patients and healthy controls were analyzed using the limma package in R. According to the screening criterion of |log FC| ≥ 1 and adjust P < 0.05, a total of 477 DEGs in the GSE25504 dataset were selected, of which 303 genes were upregulated and 174 genes were downregulated. In GSE26440, 463 DEGs were screened, including 244 upregulated and 219 downregulated genes. The DEGs of these two gene datasets are displayed using volcano plots (Figure 2A, B). Then, these DEGs were further filtered using the Venn software, and 125 overlapping genes were obtained in the

---

Figure 1. Flow chart of research design in this study. GSE25504, a discovery set, which was processed on the GPL6947 platform. GSE25504-V was a validation set and different from GSE25504, as described in “Materials and Methods”.
discovery set GSE25504 and GSE26440 (Figure 2C), among which 85 genes were upregulated and 40 genes were downregulated in sepsis samples compared to controls (Supplementary Table S2).

STRING software is an online database aimed at assessing protein-protein interaction (PPI) information. To determine the interactive profile of the 125 common genes in NS, the PPI network was constructed based on the STRING software and is shown in Figure 2D. As can be seen from the figure, genes with a higher degree of connectivity were closer to the core of the network (Figure 2D). Subsequently, an unsupervised hierarchical clustering analysis of 125 common genes was conducted in the discovery set. The horizontal axis shows neonatal samples, and the vertical axis represents the gene expression of DEGs as displayed in Figure 2E. The clustering heatmap revealed that patient samples were automatically clustered into two major groups (only one NS patient exception) under unsupervised classification (Figure 2E). As shown by the heatmap, the 125 genes displayed different expression patterns and could markedly distinguish sepsis patients from healthy individuals, which further suggested that they were closely related to NS (Figure 2E, Supplementary Figure S1). Therefore, these DEGs were selected for the screening of candidate markers for neonatal sepsis.

3.2. Gene set variation analysis of the common DEGs

To understand the changes in biological functions and signaling pathways in neonatal sepsis, we performed Gene Ontology (GO) and KEGG pathway analysis for 125 common genes via gene set variation analysis (GSVA). The results for the GO terms related to biological processes (BP) are demonstrated in Figure 3A. Exocytosis, myeloid leukocyte mediated immunity, myeloid leukocyte activation, endocytosis, and cell activation involved in immune response were significantly upregulated in NS compared to the controls, while antigen receptor mediated signaling pathway, T cell receptor signaling pathway, lymphocyte and T cell activation associated process, and adaptive immune response were suppressed in the sepsis samples (Figure 3A). Meanwhile, the upregulated KEGG pathways in sepsis patients were mainly enriched in starch and sucrose metabolism, insulin signaling pathway, and glycometabolism-related pathways (Figure 3B). In addition, the downregulated KEGG terms were primarily associated with T cell receptor signaling pathway, propanoate and pyruvate metabolism, and cysteine and methionine metabolism (Figure 3B). Analysis of gene signatures in pathways revealed that pathways clustering in a variety of biological processes and functions were distinctly different between sepsis and healthy individuals. These results demonstrated that leukocyte activation and metabolism-related processes, especially glycometabolism, were markedly upregulated in NS, whereas adaptive immune-associated biological processes were downregulated.

3.3. Immune cell infiltration profiling in sepsis patients

From the GSVA analysis, we found that immune response-related processes and pathways were significantly downregulated in neonatal sepsis. Therefore, we conducted immune cell infiltration analysis to further understand the differences in immune cell phenotypes between normal and septic neonates. Among the 22 immune cell types, follicular helper T cells were not detected in any datasets, while resting dendritic cells, activated mast cells, activated NK cells, and macrophage M1 were only detected in two datasets. As presented in Figure 4, the infiltration levels of most immune cell subtypes related to adaptive immunity, including T cell CD8, T cell CD4 naive, and resting T cell CD4 memory of GSE25504 and GSE69686 datasets, were significantly lower in the sepsis samples than in the healthy controls. The infiltration level of naïve B cells in GSE25504, as well as B cell memory and gamma delta T cells in GSE26440, was decreased in sepsis. However, only the content of regulatory T cells (Tregs) was higher in neonatal sepsis of GSE69686 (Figure 4). With regards to the infiltration levels of cell types related to innate immunity, the content of macrophage M0 and neutrophils was markedly higher in septic neonates of all three datasets. Moreover, the infiltration of active dendritic cells was higher in the sepsis group of GSE25504 and GSE69686 (Figure 4). While the NK cell content was lower in the sepsis patients in these two datasets. Taken together, these findings imply that adaptive immunity is largely suppressed in neonatal sepsis, while the innate immune response is partially activated.

3.4. Gene signature construction and identification of candidate biomarkers associated with neonatal sepsis

In order to select the genes with the best performance in the diagnosis of NS, we calculated the AUC values of common DEGs in GSE25504 and GSE26440 via ROC analysis. A total of 57 genes with AUC >0.9 in both datasets were obtained as potential candidate genes for subsequent analysis. To further narrow the scope of diagnostic markers, a LASSO regression model was applied to analyze these 57 genes in the discovery set, the regression coefficients of which are presented in Figure 5B. We employed ten-fold cross-validation with binomial deviance to select the best lambda value, which was used to construct the optimal model. As can be seen from the figure, the model performed best when 7 genes were included (Figure 5A). Therefore, seven candidate diagnostic genes for NS were identified and defined as signature genes: SLC2A3, PSTPIP2, PRRT3, OSCAR, ID3, CD3G, and CD96.

Subsequently, we validated the ability of the seven-gene signature to distinguish sepsis from uninfected individuals in independent datasets by using logistic regression combined with ROC analysis. Consistent with the results in the discovery set (the AUC value of two datasets was 1, as shown in Supplementary Figure S2), the signature genes demonstrated good performance in both the validation set of GSE25504 and GSE69686, with AUC values were 0.9524 and 0.8976, respectively (Figure 5C-D). In the meantime, we applied the ROC curve to evaluate the diagnostic implication of each gene among the 7-signature genes, and 5 genes with AUC greater than 0.75 in all datasets of this study were further screened. The results of ROC analysis of these five genes, namely SLC2A3, PSTPIP2, OSCAR, ID3, and CD3G, are plotted in Figure 5E-H. As shown in the figures, the five genes could remarkably discriminate sepsis samples and controls, which implied that they may have crucial values in the diagnosis of NS. Therefore, we defined them as candidate diagnostic biomarkers.

3.5. Expression and immune cells correlation analysis of potential markers in neonatal sepsis

To further investigate the clinical relevance of these 5 candidate genes, their relative mRNA expression levels in the discovery set and validation set were calculated and are displayed in Figure 6A. The results showed that CD3G and ID3 were significantly downregulated in NS compared to controls in the discovery set of GSE25504, GSE26440, validation set of GSE25504, and GSE69686, while OSCAR, PSTPIP2, and SLC2A3 were markedly upregulated in sepsis patients in all datasets (Figure 6A). Furthermore, the correlation between hub genes and immune infiltrating cells was analyzed using Pearson correlation coefficient. The results indicated that downregulated genes were positively correlated with most adaptive immune-related cells, such as naïve B cells, CD4 T cells, and CD8 T cells, while strongly negatively correlated with majority innate immune-related cells, such as monocytes, macrophages M0, mast cells activated, and neutrophils (Figure 6B). In contrast, the correlation between the upregulated genes and immune infiltrating cells was exactly the opposite of that of the downregulated genes (Figure 6B).

3.6. Biological functional analysis of candidate biomarkers

We conducted functional analysis to explore the underlying biological significance of the gene markers and genes that interact closely with them. Pearson’s correlation analysis was applied to assess the correlation between hub genes and all 639 differentially expressed genes in the discovery set. Based on the criterion of correlation coefficient ≥ 0.7, we
Figure 2. Identification of overlapping differentially expressed genes (DEGs) in neonatal sepsis (A, B) Volcano plots showing DEGs between control groups and sepsis patients in GSE25504 and GSE26440 datasets, respectively. DEGs were obtained in above two mRNA expression datasets. Down, downregulated genes in sepsis; None, no significant difference from controls; Up, upregulated genes in sepsis (C) Venn diagram displaying the number of overlapping DEGs in GSE25504 and GSE26440 (D) Protein–protein interaction (PPI) network of 85 upregulated genes (red nodes) and 40 downregulated genes (blue notes) were visualized by Cytoscape software. The edges between two nodes indicate the interactions of genes. The darker the node color, the higher is the connectivity of the gene (E) Heatmap representing unsupervised hierarchical clustering of the 125 overlap DEGs between sepsis patients and controls in GSE25504. Red color, upregulation in sepsis; blue, down regulation in sepsis.
Figure 3. Gene set variation analysis (GSVA) of the overlapping genes (A) Differences in biological processes of GO terms scored by GSVA between sepsis patients and controls. FC is shown from a linear model. The upper bubble and column chart represent significant upregulated biological processes in sepsis patients, and the lower bubble and column chart represent significant downregulated biological processes in sepsis patients. Top 15–16 pathway activities are shown (B) Differences in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway scored by GSVA between sepsis patients and controls. The upper bubble and column chart, up-regulated KEGG pathway in sepsis patients; the lower bubble and column chart, downregulated KEGG pathway in sepsis patients. Top 10 signaling pathways are shown. FC, fold change. NS, neonatal sepsis. P value < 0.05 was considered statistically significant.
established a gene co-expression network, in which each DEG interacted with at least one of the hub genes (Figure 7A). Because the abnormal expression of gene markers may affect the pathogenesis of sepsis by changing the expression of their co-expressed genes, we then employed enrichment analysis of biological processes and pathways for these interacting genes using Metascape. The results were visualized using Metascape and are presented in Figure 7B-D. As shown in the figures, the co-expressed genes were mainly associated with a series of biological pathways, including neutrophil degranulation, immune effector process, T cell activation, and regulation of immune response, cytokine

![Figure 4](image-url)

**Figure 4.** Immune cell infiltration profiling in sepsis patients. Comparison of infiltration levels of immune cells between neonates with sepsis and controls in three datasets GSE25504, GSE26440, and GSE69686. Green and red colors represent controls and sepsis, respectively. Data was analyzed using CIBERSORT algorithm. The levels of each immune cell type were compared using a non-parametric t-test. T, T cells. †, less than 3 samples and no statistical analysis. Statistical significance was set at P < 0.05. *p < 0.05, **p < 0.01, *** p < 0.001.
Figure 5. Identification of candidate gene markers associated with neonatal sepsis (A) A cross validation plot was generated against the log (lambda) sequence for selection of the optimal parameter (lambda) in the LASSO regression model. Two vertical dotted lines represent the two optimal values by the minimum criterion (left) and the 1. SE criterion (right). Lambda is the tuning parameter (B) LASSO coefficient profiles of the candidate genes selected by the optimal lambda. The vertical dotted line showed the optimal value (L1 Norm = 5), which was obtained from cross-validation plot. LASSO, the least absolute shrinkage and selection operator (C) ROC analysis of the gene signature for diagnosis of neonatal sepsis in GSE25504 validation set. The gene signature was identified in the LASSO regression (D) ROC analysis of the gene signature in GSE69686 (validation set) (E-H) ROC curves to show the diagnostic performance of 5 key genes in neonatal sepsis. The ROC curves of 5 gene markers in the discovery set of GSE25504 (E), in GSE26440 (F), in the validation set of GSE25504 (G), and in GSE69686 (H). AUC: area under the ROC curve.
production, and defense response (Figure 7B-D). These results suggest that upon bacterial challenge, these biomarkers may participate in inflammatory activation and immune responses by disrupting the co-expression networks, thereby playing important roles in NS development.

3.7. Validation of candidate biomarkers in clinical patients and prediction of biomarker-targeted drugs

To further validate the diagnostic performance of the five candidate biomarkers, we determined their mRNA expression levels in 22 NS patients and 16 controls using RT-qPCR. As shown in Figure 8, the expression levels of SLC2A3 and OSCAR were significantly increased in NS compared with control group. Meanwhile, the expression of CD3G in sepsis samples was significantly lower than that of controls (Figure 8). Whereas PSTPIP2 and ID3 showed no statistical difference between the two groups (Figure 6). The data indicate that three key genes SLC2A3, OSCAR, and CD3G could be novel potential biomarkers for the diagnosis of NS.

We further used DGIdb database to predict drugs that may target these biomarkers and the results were displayed in Table 1. We obtained 3 drugs targeting SLC2A3 and 10 drugs for CD3G. Unfortunately, we did not find any drugs for OSCAR. Among 10 CD3G targeted drugs, Muromonab-CD3 and Foralumab were known as its inhibitors.

3.8. Diagnostic nomogram to predict neonatal sepsis

The nomogram for predicting neonatal sepsis was developed based on the gene expression level of the screened biomarkers and is presented in Figure 9A. In this model, the largest predictor for NS was SLC2A3 (100 points), followed by OSCAR (76 points). The probability of NS occurrence corresponding to the total scores was presented in Figure 9B. A higher total score meant better discriminative ability. The ROC curve analysis was performed to evaluate the performance of this nomogram. In the discovery set, the AUC values of the ROC curve for GSE25504 and GSE26440 were 98.47% and 100%, respectively (Figure 9C-D). The AUC value of the nomogram in the validation set GSE69686 was 89.60% (Figure 9E). The results revealed that the constructed nomogram exhibited a good performance for NS prediction in our study, which further confirmed the clinical ability of these three genes as diagnostic biomarkers of NS.

4. Discussion

Neonatal sepsis (NS) is a systemic inflammatory response syndrome caused by various pathogenic bacteria and infectious factors. If not treated in time, it can lead to multiple organ dysfunction and circulatory disorders. It has been reported that in pediatric septic shock, developmental ages displayed obvious effects on the host transcriptomic
response to sepsis (Wynn et al., 2015). Regarding the pathogenesis of sepsis, newborns are not small children. Therefore, an age-specific study of the development of sepsis is necessary to improve the outcomes of NS. Due to the limitation of blood volume and quantity, research on NS is very challenging, particularly due to the lack of studies on sepsis-related molecular regulation mechanisms. In this study, we aimed to identify important gene markers related to NS using bioinformatics methods based on the online genetic database. In the discovery sets GSE25504 and GSE26440, a total of 125 DEGs were selected between sepsis patients and normal controls. In the present work, PPI networks and GSVA were used to analyze the interactions among these DEGs and the biological functions of DEGs, respectively. Furthermore, by combining the LASSO model with the ROC curves, we obtained five candidate genes strongly associated with NS from the DEGs. We further validated them in clinical samples and identified three key gene markers, which may be used as potential diagnostic markers and therapeutic targets for NS.

The GSVA on biological processes and KEGG pathways revealed that the DEGs involved in myeloid leukocyte activation, metabolism-related processes, starch and sucrose metabolism, and glycometabolism were significantly elevated in NS, whereas the DEGs associated with T cell receptor signaling pathway and adaptive immune-related pathways were significantly downregulated in neonates with sepsis. The results mean that, in response to sepsis, inflammation and metabolism-related processes, especially catabolism of carbohydrates were upregulated, while the immune response was suppressed in neonates. These findings are consistent with those of previous reports (Zeng et al., 2021a). A growing
body of evidence has shown that metabolism is crucial in the process of neonates’ immune responses to infection in sepsis (Conti et al., 2020). Smith’s study demonstrated the connections between innate immunity and metabolic pathways in neonatal sepsis (Smith et al., 2014). It is becoming increasingly clear that cellular metabolism plays a dual role in regulating immune cell function by providing sufficient biomolecules and energy for cell growth and differentiation, and directly controlling cell function using metabolic substrates, enzymes, and regulators (Loftus and Finlay, 2016; Gaber et al., 2017). These may be the reasons why metabolism-associated processes are significantly upregulated in NS.

The significant increase in glycometabolism may be explained by the following aspects. The energy storage of neonates is lower than that of adults; therefore, they have a higher basal metabolism and require more energy to sustain growth and development. Especially in the first few days, newborns start to produce glucose by themselves, replacing the supply of maternal glucose, a process that requires catabolism of carbohydrates to accelerate the production of glucose to supply organ development (Harbeson et al., 2018). Innate immune cells preferentially apply glycolysis to generate sufficient ATP to fight bacterial invasion by augmenting glucose flow and producing lactate by glycolysis (Nolt et al., 2018). Additionally, compared with the monocytes of healthy individuals, the monocytes of sepsis patients showed obviously upregulated glycolysis (Schenz et al., 2020).

Our results revealed that propanoate and pyruvate metabolism was significantly down-regulated in sepsis patients. Propanoate metabolism eventually enters the tricarboxylic acid (TCA) cycle or gluconeogenesis process. Pyruvate is the product of the glycolysis of glucose. Under aerobic conditions, pyruvate undergoes oxidative decarboxylation converted to acetyl-CoA and then enters the TCA cycle. Glucose is completely oxidized via the TCA cycle, which supplies energy to the body, and stores energy in the form of ATP (Wasylik and Zwolak, 2021). TCA cycle is the junction of the metabolism of three major macronutrients including glucose, lipid, and protein. The attenuation of propanoate and pyruvate metabolism suggests an inhibition of the TCA cycle, which accompanies with the failure of energy supply and the disorders of the metabolism of all macronutrients. Thus, sepsis-induced dysfunction and cell metabolic disorders occur in the course of sepsis, which aggravates infection-related organ damage (Wasylik and Zwolak, 2021; Zeng et al., 2021b). Consistent with the result of a previous report (Tong et al., 2022), we also found that cysteine and methionine metabolism was decreased in NS, which indicated that the metabolism of amino acids was altered in patients with sepsis.

Many previous studies have shown that in sepsis, the invading bacteria activate the innate immune system, leading to an early stage of pro-inflammatory response in the host, followed by an anti-inflammatory response with immunosuppression (Cecconi et al., 2018; Vachharajani and McCall, 2019). However, a recent publication reported that the anti-inflammatory process was activated quickly during the initial stage of sepsis (Hotchkiss et al., 2009). The dysregulated innate immunity and restrained adaptive immunity caused by sepsis together trigger persistent pro-inflammatory and anti-inflammatory pathways, which eventually

![Figure 8](image_url)  
Figure 8. The expression of 5 key genes between neonatal sepsis group and control group. mRNA expression levels of SLC2A3, PSTPIP2, OSCAR, ID3, and CD3G in septic neonates (n = 22) and controls (n = 16) were determined by RT-qPCR, using GAPDH as housekeeping gene. *p < 0.05, **p < 0.01.

### Table 1. Prediction of drugs targeting gene markers.

| Gene      | Drug               | Drug Class      | Sources              | Sources              |
|-----------|--------------------|-----------------|----------------------|----------------------|
| SLC2A3    | GLUFOSFAMIDE       | Small molecule  | TdgClinicalTrial     | ChemblInteractions   |
|           | RESVERATROL        | Small molecule  | DTC                  |                      |
|           | INSULIN            | Protein         | NCI                  |                      |
| CD3G      | MUROMONAB-CD3      | Monoclonal antibody | TdgClinicalTrial  | ChemblInteractions   |
|           | MEDI-565           | n/a             | TTD                  |                      |
|           | FORALUMAB          | Monoclonal antibody | TdgClinicalTrial  | ChemblInteractions   |
|           | APM-11             | n/a             | TTD                  |                      |
|           | TEPLIZUMAB         | Monoclonal antibody | TTD                |                      |
|           | MGD-007            | n/a             | TTD                  |                      |
|           | OTELIXIZUMAB       | Monoclonal antibody | TdgClinicalTrial  |                      |
|           | VISILIZUMAB        | Monoclonal antibody | TdgClinicalTrial  |                      |
|           | ERTUMAXOMAB        | n/a             | TTD                  |                      |
|           | TEBENTAFUSP        | n/a             | TTD                  |                      |
results in tissue and organ dysfunction, while immunosuppression with T cell exhaustion accelerates sepsis-related mortality (Delano and Ward, 2016; Patil et al., 2016).

Based on the transcriptomic data, we analyzed the characteristics of infiltrating immune cells in healthy and septic neonates using CIBERSORT. Our results demonstrated that the infiltration levels of some innate immune cells, including macrophage M0, neutrophils, and dendritic cells, were prominently higher in sepsis samples. Once the host is invaded by pathogenic bacteria, the innate immune system, which includes monocytes, neutrophils, and macrophages, is activated, releasing large amounts of pro-inflammatory cytokines and chemokines, which can cause a persistent inflammatory response (Hotchkiss et al., 2013). Neutrophils are the most important component of innate immunity and are crucial for controlling early pathogen invasion (Tamayo et al., 2012). The study on NK cells was not thorough enough in human sepsis. We showed that the content of NK cells decreased in NS. This is consistent with a previous study, which reported that the number of NK cells in the circulation was reduced in sepsis (Venet et al., 2010). The lower the number of NK cells, the higher the sepsis-related mortality (Giamarellos-Bourboulis et al., 2006). In addition, consistent with the results of GSVA enrichment, the infiltration levels of cells involved in adaptive immunity were significantly decreased in the present study. This implied that, along with innate immune response, an anti-inflammatory response with immunosuppression was also prevalent in neonatal sepsis. Interestingly, the results of previous reports in pediatric sepsis were similar to our findings: the expression of genes associated with innate response was

Figure 9. The nomogram model for the diagnosis of neonatal sepsis (A) The nomogram prediction score of neonatal sepsis based on the screened biomarkers (B) The incidence risk of NS corresponding to the total points (C–E) The ROC curve for assessing the discriminative ability of the nomogram in the discovery set and validation set.
upregulated and genes associated with acquired immunity were suppressed (Wong, 2013).

In this study, we acquired a 7-gene signature using ROC analysis and LASSO model. Then, the AUC value was further used to validate the diagnostic performance of each gene among the gene signature in two validation datasets and clinical patients, and 3 hub genes, SLC2A3, OSCAR, and CD3G, were identified as important biomarkers for NS. SLC2A3 (Solute carrier family 2 member 3), also known as glucose transporter 3 (GLUT3), is low or even not expressed in many tissues, except nerves and neural tissue (Masin et al., 2014). Fu et al. identified SLC2A3 as a hub gene for sepsis development from three datasets via bioinformatics analysis, and they also revealed that the expression of SLC2A3 was significantly increased in endothelial cells (ECs) upon LPS stimulation (Fu et al., 2020). LPS-induced endothelial inflammation plays a vital role in the progression of sepsis. In addition, a previous study has shown that increased levels of GLU3 expression act as an energy source for infiltrated macrophages, while the inhibition of glucose uptake via the suppression of GLU3 expression prevents LPS-mediated inflammation in macrophages (Reddy et al., 2010). Therefore, SLC2A3 may be involved in microbial sepsis and endotoxemia and is expected to become a potential therapeutic option for NS.

Osteoclast-associated receptor (OSCAR) is an active regulator of osteoclastogenesis, whose main function is to regulate bone development and maintenance (Barrow et al., 2011). Besides osteoblasts, OSCAR is expressed in other human myeloid cells, such as dendritic cells, monocytes, neutrophils, and macrophages (Merck et al., 2006). The interaction of collagen and OSCAR can facilitate dendritic cell maturation and promote the release of pro-inflammatory cytokines and chemokines including TNF-α, IL-6, IL-8, and IL-10 (Merck et al., 2005). Recent studies have suggested that inflammatory molecules induce high expression levels of OSCAR in ECs and macrophages (Sinningen et al., 2013). Although no studies have indicated a link between OSCAR and sepsis, considering that OSCAR has a role in the activation of most types of immune cells and ECs, we speculate that OSCAR may participate in inflammation-related diseases and is perhaps also very important in the development of sepsis. A recent study established that the expression of CD3G (T-cell surface glycoprotein CD3 gamma chain) was inversely associated with sepsis severity and mortality (Almansa et al., 2015). In our work, we found that CD3G was prominently reduced in neonates with sepsis compared to healthy controls. Thus, we speculate that CD3G may be a potential marker for NS diagnosis.

Compared to the existing diagnostic gene signatures, our three gene markers showed a better discrimination performance for NS than Septi-Cyte Lab (McHugh et al., 2015), and were more concise and convenient than PEDSEPS-GBM, which is an eighteen gene combination model with a high predictive specificity for sepsis in children (Ying et al., 2021). In our study, three biomarkers SLC2A3, OSCAR and CD3G could effectively and accurately predict NS either alone or in combination. With the development of prospective studies, we believe that this gene panel will provide help for the clinical diagnosis of NS.

We should consider that our study is not without limitations when interpreting our data. First, the datasets of neonatal sepsis in the public databases are very scarce and we only screened 3 sepsis datasets containing neonates, with a small sample size. Thus, studies with larger sample sizes of NS are needed to validate the value of these biomarkers in early diagnosis and therapy. Second, because sepsis datasets are scarce in neonates, sepsis and septic shock were not distinguished in our study, and septic shock was included in sepsis for data analysis. Moreover, the biological roles and molecular mechanisms of the gene markers need to be further investigated in the future. Even so, the results presented in this study are meaningful, because neonatal samples are not easily available, and only a few reports have investigated the regulatory mechanisms of this disease. Thus, our study may make up for this deficiency and lay the foundation for subsequent research on NS.

5. Conclusions

In summary, our results indicate that metabolic processes and inflammatory responses are upregulated in NS, while immune responses are suppressed. Through a series of bioinformatics analyses and clinical validation, we identified three promising markers for NS, which could specifically and accurately differentiate sepsis patients from healthy individuals. These biomarkers not only have great potential in the early diagnosis and intervention of NS, but may also provide novel insights into NS pathogenesis.

Declarations

Author contribution statement

Rui Yan: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Tao Zhou, Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

Rui Yan was supported by Basic and Applied Basic Research Foundation of Guangdong Province [2019A1515110548], National Natural Science Foundation of China [82101804].

Data availability statement

Data included in article/supp. material/referenced in article.

Declaration of interest’s statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e11634.

Acknowledgements

We would like to thank Baowen Zhuo and Feng Wang for their excellent help in data collection and analysis.

References

Almansa, R., Heredia-Rodríguez, M., Gómez-Sanchez, E., Andaluz-Ojeda, D., Iglesias, V., Rico, L., et al., 2015. Transcriptomic correlates of organ failure extent in sepsis. J. Infect. Dis. 15 (12), e854–e858.

Altunhan, H., Annagir, A., Oys, R., Mehmetoğlu, Í., 2011. Procalcitonin measurement at 24 hours of age may be helpful in the prompt diagnosis of early-onset neonatal sepsis. Int. J. Infect. Dis. 15 (12), 234–242.

Barrow, A.D., Raynal, N., Anderlini, T.L., Slatter, D.A., Bihan, D., Pugh, N., et al., 2011. OSCAR is a collagen receptor that contributes osteoclastogenesis in DAP12-deficient humans and mice. J. Clin. Invest. 121 (9), 3505–3516.

Cecconi, M., Evans, L., Levy, M., Rhodes, A., 2018. Sepsis and septic shock. Lancet 392 (10141), 75–87.

Chauhan, N., Tiwari, S., Jain, U., 2017. Potential biomarkers for effective screening of neonatal sepsis infections: an overview. Microb. Pathog. 107, 234–242.

Chen, H., Li, Y., Li, T., Sun, H., Tan, C., Gao, M., et al., 2019. Identification of potential transcriptional biomarkers differently expressed in both S. Aureus- and E. Coli-induced sepsis via integrated analysis. BioMed Res. Int. 2019, 2487921.

Conti, M.G., Angelidou, A., Diray-Aree, J., Smolen, K.K., Lasky-Su, J., De Curtis, M., et al., 2020. Immunometabolic approaches to prevent, detect, and treat neonatal sepsis. Pediatr. Res. 87 (2), 399–405.

Delano, M.J., Ward, P.A., 2016. The immune system’s role in sepsis progression, resolution, and long-term outcome. Immunol. Rev. 274 (1), 330–353.

Fleischmann-Struzewicz, C., Goldfarb, D.M., Schlattmann, P., Schlaphbach, L.J., Reinhart, K., Kissoon, N., 2018. The global burden of paediatric and neonatal sepsis: a systematic review. Lancet Respir. Med. 6 (1), 223–230.

