Weighted Gene Co-Expression Network Analysis and Machine learning Identify Critical Genes in the Development of Osteomyelitis after Staphylococcus aureus Infections

Baixing Chen  
KU Leuven

Shaoshuo Li  
Nanjing University of Chinese Medicine

Zhaoqi Lu  
The Fourth Clinical Medical College, Guangzhou University of Chinese Medicine, China

Mingling Huang  
Guangzhou University of Chinese Medicine

Shi Lin  
Guangzhou University of Chinese Medicine

Hang Dong  
The First Affiliated Hospital of Guangzhou University of Chinese Medicine

Research Article

Keywords: Osteomyelitis, bone infections, WGCNA, machine learning, random forest, development, Staphylococcus aureus, GEO, recursive feature elimination, microarray

DOI: https://doi.org/10.21203/rs.3.rs-440346/v1

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

**Background:** Staphylococcus aureus (S. aureus) is the most common pathogen that causes osteomyelitis (OM). However, OM's pathogenesis, which is not clear, involves many factors such as environment, genetics and immunity dysregulation. This study aims to explore the key genes involved in the pathogenesis and development of OM following S. aureus infection.

**Methods:** After obtaining the datasets of GSE6269 and GSE16129, we performed weighted gene co-expression network analysis (WGCNA) to find clusters modules of highly correlated genes and recursive feature elimination (RFE) method to narrow the range of feature genes. For determining the effect of feature genes, we constructed a random forest (RF) model with feature genes and validated the predictive validity of the RF model using independent data from GSE11908. The protein-protein interaction (PPI) network identifies essential proteins that contributed to OM development.

**Results:** There were 12,401 genes from 77 samples that 48 S. aureus patients developed to OM and 29 of those without OM. We divided 31 significant gene modules into different modules, and the brown module significantly related to OM. Biological Functions of the brown module mainly enriched in the inflammatory response, metabolic, cancer, viral pathways, protein binding and RNA binding. After screening, 19 genes, including CYP2E1, BBS10, ARPC5L, GAPVD1, PURA, RBMS1, BTN2A2, EXOSC8, METTL8, FYCO1, KHK, PRPF38B, CD72, C2CD5, ABHD6, CD200, FAM53C, HCP5 and ELP1, were defined as feature genes for constructing RF model. After validating the external data, the average area under the curve was 85%, and the accuracy of the RF model was 85.7%. The protein function of modules enriched in the RNA exosome complex's catalytic component and regulation of actin polymerization.

**Conclusions:** This study aimed to identify related genes involved in the occurrence and development of OM. We constructed the RF model with 19 genes, which effectively classify the patients with OM or non-OM. Despite its limitations, the study certainly adds to our understanding of OM's pathogenesis, and therefore, has significant implications for potential therapeutic targets and the predicted value of OM.

Introduction

Osteomyelitis (OM) is the commonest musculoskeletal infection in children, which results from the bacteria travel through blood flow to the epiphyseal plate of long bones. The estimated annual incidence of osteomyelitis is more than 10 per 100,000 children [1-4]. The OM cannot be underestimated, because it causes multiple complications with catastrophic outcomes. Moreover, Staphylococcus aureus (S. aureus) to the OM is the commonest pathogen, which will be found in almost 30% to 63% of confirmed OM cases [1, 5]. While the OM is an infectious disease infected with bacteria, a few factors increase the likelihood of OM occurring after infection. For example, Guillon et al. reported that boys and toddlers have a higher prevalence of pediatric bone infections. [2]. Also, OM may occur in patients with immunodeficiency [6, 7] and hemoglobinopathies [8], these factors are not generalizable in the majority of cases. Currently, it is acknowledged that anatomic susceptibility in the well-vascularized metaphysis to bacterial invasion is
likely pathogenesis. Although we can reveal the anatomical pathogenesis, it cannot help clinicians to diagnose and predict definitively as soon as possible. Because the initial symptoms of OM are without specificity, these symptoms may be considered as fever or upper respiratory infection. [9]. The initial occurrence of musculoskeletal symptoms and signs means that extent of infection in the bone has been exacerbated. However, the diagnostic approaches are not sensitive, and OM's pathogenesis following S. aureus is not clear. Therefore, developing early reliable predictive methods and investigating OM's related pathogenesis following S. aureus remains vital to the treatment.

High-throughput technology is a practical approach to systematically identify and screen the most potential biomarkers, that help researchers explore the fundamental biological principle. It has demonstrated that biological functions play a crucial role in the development of OM following S. aureus infection. [10]. The previous research compared the differentially expressed genes (DEGs) [11] between patients with OM and non-OM, but they did not implement the cluster analysis of largely correlated genes, that overlook the feature genes which are without differences. It still does not know the function of non-DEGs in the development of OM. Therefore, it is crucial to adopt new approaches in the study of OM.

Weighted gene co-expression network analysis (WGCNA), which is broadly acknowledged as a communication bridge between feature genes and clinical characteristics, is a systematically biological application [12]. The WGCNA method is used to screen the genes with similar expression patterns to cluster them as a module [13, 14]. Apart from WGCNA, machine learnings, such as Support Vector Machine (SVM) [15], Random Forest (RF) [16] and K-nearest neighbor (KNN) [17], have been successfully used to select specific and sensitive feature genes from a massive amount of information. Therefore, we will explore the critical genes involved in OM's pathogenesis following S. aureus infection by new approaches. This review aims to provide clues in developing therapeutic targets and potential biomarkers for the osteomyelitis.

**Materials And Methods**

**Data resource and data preprocessing**

The GSE6269 [18] and the GSE16129 [19] gene expression profile data were from the same institute tested by the GPL96 platform. The analysis contained 77 participants’ peripheral blood samples, including 48 S. aureus patients developed to OM (OM group) and 29 of those without OM (SI group). We transform the raw microarray data (TXT files) by log2 transformation (Limma 3.12) [20] and normalize the data using the median normalization method [20].

**WGCNA analysis**

The highly related modules from the co-expression network were identified using the WGCNA function from R [12, 15], whose analysis procedure was involved topological properties calculating, network construction, clustering analysis, module detection and visualization. The dynamic trees cut algorithm split genes into modules whose minimum size is 50 in every module. The significance of the module was
defined as the average absolute gene significance. It could quantify the correlation strength between clinical trait and module eigengene (ME), and calculate the regression between gene expression and clinical information. After identifying the modules, the most significantly associated module, as the key module, was selected for further analysis.

**Enrichment analysis of gene**

To understand the biological meaning of the key module. The data were uploaded into DAVID (version 6.8) [21] for Gene ontology function (GO) [22] annotation and Kyoto encyclopedia of genes and genomes pathway enrichment analysis (KEGG) [23]. The enrichment analysis would be presented when their cut-off criteria were P<0.05 and a count >5.

**Feature selection of candidate feature genes**

The significant module genes were screened out using the Recursive Feature Elimination (RFE) algorithm [24]. After removing a pile of weaker characteristics, this feature selection approach would obtain a fitting model when reached the specified number of characteristics. It is helpful to narrow the range of genes from the significant module. This method could find out feature genes as clinical biomarkers, which could effectively classify and identify the patients with OM and non-OM.

**Construction of Random Forest model**

The gene expression datasets usually contain many different noises, and random forest (RF) has good anti-noise ability. Because it introduces randomness, it is better at control over-fitting [25, 26] to avoid false variables. RF uses a random method to build a forest composed of many decision trees, and there is no correlation between each decision tree in the random forest. To verify how accurate the feature genes would be and classify samples, a random forest (RF) model was constructed with feature genes. To improve the accuracy of the model, we use the five-fold cross-validation (CV) technique. This technique randomly divided the feature genes into five subsets, four of which were training datasets, and the remaining one was a test dataset. After repeating the CV 5 times, all subsets were calculated. We would evaluate the model's effectiveness based on the receiver operating characteristic curve (ROC) and Area under the Curve of ROC (AUC).

**Validation of independent data**

The GSE11908 [27] dataset tested by the GPL96 platform was defined as an external validation set. These data were from patients with *S. aureus*, including six samples with OM and eight samples without OM. The data from these 14 samples were employed to validate the RF model.

**PPI network and module analysis**

The protein-protein interaction (PPI) network identified important proteins that are susceptible to OM development. We downloaded the interactive relationship of proteins of feature genes from the STRING
database [28]. Subsequently, the MCODE [29] function of Cytoscape was employed to analyze modules with the following parameters: node score cutoff = 0.2, thresholds of degree cutoff = 2, k-core ≥ 2 and max. depth = 100 [29].

**Statistical analysis**

The R software (version 3.6.3) was adopted to conduct all the statistical analyses in the present study. The weighted correlation network analysis was performed using the WGCNA [12] packages (version 1.61), and the RFE function from the Caret package was employed to calculate feature selection [30]. The random-forest classifier from the random forest package [31] was used for classification analysis, and ROC analyses were performed using the ROCR package when $P$-value < 0.05 were considered to indicate a statistically significant difference.

**Results**

**Gene Co-expression Networks of Osteomyelitis**

We adopted the WGCNA function to screen the OM-related gene from GSE6269 and GSE16129. Firstly, no outliers were detected by the Hclust function, and 12401 candidate genes were involved in the subsequent analysis. When a power of $\beta = 5$ was selected as the soft-threshold to ensure a scale-free network, it was implemented to measure the co-expression network (Figure 1). Thirty-one significant gene modules (Figure 2) then were assigned to different colors. After calculating each module by WGCNA, the correlations between gene expression and osteomyelitis (Figure 3) significantly demonstrated in the brown module ($R^2 = 0.57$, $P<0.001$), which was defined as candidate genes (Figure 4). There were 918 genes in the brown module for the following analyses.

**Biological Functions and Pathways Involving the Candidate Module**

The GO enrichment analysis of the candidate genes showed that the cellular component (CC) of genes concentrated on the cell nucleus and particular extracellular region part. Besides, the candidate genes were mainly involved in three biological processes (BP), including protein binding, poly (A) RNA binding and ATP binding. The top three of molecular function (MF) were RNA polymerase II promoter, regulation of transcription from RNA polymerase II promoter and negative regulation of transcription (Figure 5). For the KEGG pathways, the candidate genes were significantly related to pathways including metabolic pathways, pathways in cancer, human papillomavirus infection, RNA transport, thermogenesis, PI3K-Akt signaling pathway, Epstein-Barr virus infection, Human T-cell leukemia virus 1 infection, FoxO signaling pathway and AMPK signaling pathway (Figure 6).

**Recursive feature elimination (RFE) for Identification of feature genes**

To narrow the range of candidate genes, we adopted the RFE algorithm to select the feature genes. The prediction accuracy was the highest (81%) when selected 19 genes from the brown module (Figure 7).
These genes were involved as the feature genes to construct the Random forest model. These feature genes included Cytochrome P450 2E1 (CYP2E1), Bardet–Biedl syndrome 10 (BBS10), Actin Related Protein 2/3 Complex Subunit 5 Like (ARPC5L), GTPase Activating Protein And VPS9 Domains 1 (GAPVD1), Purine Rich Element Binding Protein A (PURA), RNA Binding Motif Single Stranded Interacting Protein 1 (RBMS1), Butyrophilin subfamily 2 member A2 (BTN2A2), Exosome component 8 (EXOSC8), Methyltransferase Like 8 (METTL8), FYVE And Coiled-Coil Domain Autophagy Adaptor 1 (FYCO1), Ketohexokinase (KHK), Pre-mRNA Processing Factor 38B (PRPF38B), Cluster of Differentiation 72 (CD72), C2 Calcium Dependent Domain Containing 5 (C2CD5), Alpha/beta-Hydrolase domain containing 6 (ABHD6), Cluster of Differentiation 200 (CD200), Family With Sequence Similarity 53 Member C (FAM53C), HLA Complex P5 (HCP5) and Elongator Complex Protein 1 (ELP1).

The construction of the classification model by the random forest algorithm

To validate the accuracy of feature genes, the feature genes were involved in a random forest algorithm for constructing classification models. After repeating the CV 5 times, the mean value of AUC was 0.85 ± 0.09 in the ROC curve (Figure 8), and the average accuracy of the model was 85.7% (95%CI: 0.7587-0.9265). The external dataset validated the efficiency of the random forest model. The value of AUC was 0.82 (Figure 9), and the accuracy of external validation was 78.6%. Besides that, the positive predictive value was 100%, and the negative predictive value was 66.7% that the confusion matrix visualized these results (Figure 10).

Hub modules screening of the PPI network

The PPI network of feature genes, with 38 nodes and 89 edges, was established by the STRING website and the Cytoscape. The top ten proteins were ACTR3, ARPC1A, ARPC2, ARPC3, SKIV2L2, ARPC5L, EXOSC5, EXOSC7, EXOSC8 and GAPVD1 (Figure 11). To screen the cluster of proteins, we selected two modules from these proteins using MCODE. The protein function of modules enriched in the catalytic component of the RNA exosome complex and regulation of actin polymerization, that these two modules were defined as module1 and module 2 respectively. The 4 nodes, Superkiller viralicidic activity 2-like 2 (SKIV2L2), Exosome component 5 (EXOSC5), Exosome component 7 (EXOSC7), Exosome component 8 (EXOSC8), belonged to module 1. Five other nodes in module 2 were Actin Related Protein 3 (ACTR3), Actin Related Protein 2/3 Complex Subunit 1A (ARPC1A), Actin-related protein 2/3 complex subunit 2 (ARPC2), Actin Related Protein 2/3 Complex Subunit 3 (ARPC3), Actin-related protein 2/3 complex subunit 5-like protein (ARPC5L) and GTPase Activating Protein And VPS9 Domains 1 (GAPVD1).

Discussion

In this research, we merge two GEO data as one dataset to find the specific modules using WGCNA, and the genes from the brown module were detected as candidate genes. After reducing the number of useless candidate genes using the RFE, 19 feature genes construct the RF model, whose mean value of AUC was 0.85 ± 0.09 and average accuracy of the model was 85.7%. After external verification for the random forest model to evaluate the predictive power, the analysis showed that the value of AUC was
0.82, and the accuracy of external validation was 78.6%. Consequently, this study demonstrated that these feature genes could be functional biomarkers to identify patients with OM after S. aureus infection. To our knowledge, this is the first study to identify feature genes in OM development by WGCNA and machine learning.

Some published studies suggested that the host blood transcriptional profiles play an essential role in diagnosing infectious diseases [18, 32-34]. It is worth noting that one previous study [11] divided samples into three comparison groups, including all of the S. aureus patients vs. Ctrl group (i.e., healthy people), OM patient vs. Ctrl group, and OM patients vs. non-OM patient following S. aureus infection. They defined the genes as the DEGs when they overlap the three comparison groups. This method is an interesting approach to exploring the co-expression genes from every group, but it is hard to explain the effect of these genes in the development of OM following S. aureus infection due to the indirect comparison of OM and non-OM patients following S. aureus infection. Besides, this method overlooked the importance of genes that are without statistical significance.

Furthermore, Wang et al. [35] employed gene expression profile GSE16129 for analyses and screened the DEGs with the approach as previously reported by Chen et al. [11]. Although they verified the candidate genes by mice models of implant-associated S. aureus osteomyelitis, it should note that patients are children who were encountering hematogenous osteomyelitis but not traumatic osteomyelitis. However, their study used the animal model with traumatic osteomyelitis that could not demonstrate the genes from patients with hematogenous osteomyelitis. Our study identified the gene module related to OM after S. aureus infection and determined the feature genes as early biomarkers by novel approaches.

In our study, we found that the brown module enriches pathways in cancer and metabolic pathways. This finding is consistent with those of Chen et al. [11]. One unanticipated discovery was that viral pathways, such as human papillomavirus infection, Epstein-Barr virus infection and Human T-cell leukemia virus 1 infection pathways, were related to the genes of OM. There are similarities between the attitudes expressed in this study and those described by Droz et al. [36]. Droz et al. [36] reported significant differences in the amount of human rhinovirus between patients with Kingella kingae osteoarticular (KKO) and non-KKO.

Similarly, Basmaci et al. [37] suggested that the respiratory virus is highly associated with the pathophysiology of Kingella kingae osteoarticular. In previous studies, the interaction between viral and bacterial was related to the pathogenesis of infectious diseases [38], and the presence of the virus may increase the susceptibility to bacterial colonization. It has been demonstrated that Streptococcus pneumonia easily adheres and invades after influenza virus infection [39]. Likewise, the Respiratory syncytial virus plays a crucial role in the development of S pneumonia infection [40]. This observation may support the hypothesis that the patients infected with the virus and S. aureus are possibly easier to develop into OM. Therefore, the role of viral infections in the pathophysiology of OM is worth emphasizing and lucubrating. The PI3K-Akt signaling pathway [41], FoxO signaling pathway [42] and
AMPK signaling pathway [43], which are playing a crucial role in a wide range of key cellular functions, are the popular subject investigated which mainly concern immune response.

Two highly interconnected regions in a network were found from 19 feature genes. The exosome complex, a multi-protein intracellular complex, can degrade RNA and transfer multiple cargos to mediate intercellular communication [44]. Previous research demonstrated that the exosome complex plays an important role in musculoskeletal diseases, including muscle regeneration, age-related bone loss, and osteoarthritis [45]. These findings supported that there are closer connections between the RNA exosome complex and the development of OM. These findings support the conceptual premise that such connections exist between RNA exosome complex and progression of OM. Besides, regulation of the actin polymerization is related to actin-associated proteins that participate in a diverse array of cellular processes. Holliday et al. showed that the exosome from osteoclast selectively collaborated with actin-associated proteins involved in the regulation signals of bone remodeling [46]. These targets and mechanisms could be the starting point for the research of OM therapy. After external data validation, the RF model showed great classification capacity and could be applied to classify the high-risk patients who are easier to develop into OM after infecting S. aureus.

The main weakness of this study was the lack of literature, that can support the effect of 19 feature genes in the development of OM. More genetic information about OM would help us to establish a greater degree of accuracy on this matter in the future. An additional uncontrolled factor is that computing power is insufficient to calculate many genes with cross-validation repeatedly. It may affect the accuracy of the prediction model. Besides, the sample size of external data is small, that it cannot sufficiently verify the efficacy of the RF model.

Conclusion

In conclusion, this study set out to identify feature genes involved in the development of OM following S.aureus using WGCNA and RFE. Nineteen feature genes, including CYP2E1, BBS10, ARPC5L, GAPVD1, PURA, RBMS1, BTN2A2, EXOSC8, METTL8, FYCO1, KHK, PRPF38B, CD72, C2CD5, ABHD6, CD200, FAM53C, HCP5, and ELP1, were used to construct the RF model to classify the high-risk patients. Despite its limitations, the study certainly adds to our understanding of the pathogenesis of OM following S. aureus infection and may, therefore, have significant implications for potential therapeutic targets of OM.

Abbreviations

OM: osteomyelitis; WGCNA: weighted gene co-expression network analysis; RFE: recursive feature elimination; RF: random forest; PPI: protein-protein interaction; DEGs: differentially expressed genes; SVM: Support Vector Machine; KNN: K-nearest neighbor; ME: module eigengene; GO: Gene ontology function; KEGG: Kyoto encyclopedia of genes and genomes pathway enrichment analysis; RFE: Recursive Feature Elimination; CV: cross-validation; ROC: receiver operating characteristic curve; AUC: Area under the Curve of ROC; CC: cellular component; BP: biological processes; MF: molecular function
Declarations

Acknowledgements

We thank members of the Guangzhou University of Chinese Medicine lab for their thoughtful comments.

Authors’ contributions

HD revised the final manuscript and was responsible for the funding. BHC designed this research and drafted the manuscript. He also searched the databases, analysed the data and created the illustrations. SSL provided valuable suggestions for data analysis and figures design. ZQL, MLH and SL were involved in the design of the manuscript. All authors reviewed the manuscript.

Ethics approval and consent to participate

Not applicable.

Availability of data and materials

This article does not contain any studies with human participants or animals performed by any of the authors. Because publicly available datasets were analyzed in this study. This data can be found here: The datasets of GSE6269, GSE16129 and GSE11908 were downloaded from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing financial interests.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 82004390).

References


**Figures**
Analysis of the scale-free network coefficient R-squared for the soft threshold ($\beta$) and the mean connectivity for the soft threshold. We hope that the value of R-squared exceeds 0.85, so the power value is 5 ($\beta = 5$).
Figure 2

Building a cluster dendrogram based on the dissimilarity of the topological overlap, which presents 31 network modules.
Figure 3

Heatmap demonstrating the correlation between module eigengenes and clinical characteristic
Figure 4

Correlation analysis of the genes in module brown
Figure 5

GO enrichment analysis of brown module.
Figure 6

KEGG pathways analysis of brown module.
Figure 7

Feature genes selected using RFE analysis. The x-axis is the feature gene number, and the y-axis indicates the corresponding prediction precision. The gene combination with the highest precision is marked in the blue spot, which was a 19-gene combination.
Figure 8

ROC curve of the Random forest. The x-axis represents the false positive rate, and the y-axis indicates the true positive rate. The pink line marks the simulation results (mean ROC).
Figure 9

ROC curve of the Random forest verified by the GSE11908 dataset. The x-axis represents the false positive rate, and the y-axis indicates the true positive rate. AUC, the area under the curve.
Figure 10

Confusion matrix of the random forest classifier.
Figure 11

The top ten proteins in the 19-gene combination