Machine Learning and Bioinformatics Approaches to Identify the Candidate Biomarkers in Severe Asthma

Fuying Zhang  
Zhangjiajie Hospital Affiliated to Hunan Normal University

jiabao zhu  
The Second Affiliated Hospital of Nanchang University

Mingsheng Lei (✉ mingshenglei@163.com)  
Zhangjiajie Hospital Affiliated to Hunan Normal University

Research Article

Keywords: severe asthma, machine learning, inflammation, candidate biomarkers, glucocorticoid

Posted Date: March 31st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2724025/v1

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

Background

Severe asthma is defined as a persistent increase in airway inflammation despite the use of systemic glucocorticoids, targeted biologic therapies. Early prediction of severe asthma is challenging due to the lack of valuable biomarkers. The aim of this study was to identify crucial differentially expressed genes (DEGs) associated with severe asthma through approaches of bioinformatics analysis.

Methods

Three datasets GSE130499, GSE43142 and GSE43696 were derived from the GEO expression database. Two datasets (GSE130499 and GSE43142) were merged, and batch effects were removed by using the "SVA" package. Afterwards, the differentially expressed genes (DEGs) were analyzed with the "limma" package. Next, DEGs were functionally enriched and pathway analyzed in the online analysis website DAVID, then DEGs were analyzed again by two machine learning algorithms (LASSO and SVM-RFE) to obtain the candidate biomarkers, and the diagnostic validity of the biomarkers was assessed using subject operating characteristic (ROC) curves, and finally the results were further validated through the GSE43696.

Results

Total of 73 gene differential expression genes were identified in severe asthma and normal control. After screening with two machine learning algorithms, LASSO and SVM-RFE, three genes (BCL3, DDIT4 and S100A14) were recognized as biomarkers of asthma and had good diagnostic effect. Among them, BCL3 transcript level was down-regulated in severe asthma, while S100A14 and DDIT4 transcript levels were up-regulated.

Conclusions

In this study, we identified three differentially expressed genes (BCL3, DDIT4 and S100A14) of diagnostic significance that may be involved in the development of severe asthma and proposed new insights into the underlying mechanisms.

1 Background

Asthma is a complex, heterogeneous, chronic inflammatory airway disease that affects more than 300 million people worldwide (1). Asthma is related to chronic airway inflammation, reversible expiratory airflow limitation and airway hyperresponsiveness, causing different symptoms such as shortness of breath, cough and chest tightness (2). The biological pathways of severe asthma are generally classified...
as type 2 hyperinflammatory (2-high) and type 2 low-inflammation (2-low) according to the infiltration of various immune cells (eosinophils, neutrophils) or exhaled biomarkers (elevated nitric oxide (Feno)) (3). Most patients with asthma have disease control with inhaled corticosteroids (ICS) combined with or without long-acting beta2 agonists (LABA), and some severe patients have disease control with biological targets that specifically block the T2 pathway (3), However, the symptoms are hardly controlled among some severe asthma individuals, even treated with systemic steroids or biologic therapy (1, 4), these patients may have multiple therapeutic targets and altering one pathway in the complex asthma pathophysiologic setting may not completely control asthma (5). The consequences of aggravated asthma are greater loss of lung function, decreased quality of life, and increased risk of hospitalization and death (6, 7), therefore it is essential to monitor the progression of asthma to severe disease; however, there are no validated candidate genes to predict severe asthma or prevent the progression of mild to moderate asthma.

The integrated data analysis and network-based approaches can help identify clinically useful biomarkers (8). Machine learning has significantly improved the predictive and accuracy value of key genes identified based on microarrays and next-generation sequencing data (9). It analyzes large amounts of data and establishes complex nonlinear relationships in order to produce the desired results (10). Three learning strategies are included in machine learning, including supervised learning, semi-supervised learning, and unsupervised learning. In this study, we utilized supervised learning due to the purpose of supervised learning is to predict data, which consists of fitting a model with labeled training data and then using it to make predictions, which can be categorized as either regression (where the predictor variable is numerical) or classification (where the predictor variable is categorical) problems. Machine learning methods for supervised learning strategies include Artificial Neural Networks (ANNs), Bayesian Networks (BNs), The least absolute shrinkage and selection operator (LASSO) regression, etc. (11, 12) Using machine learning to help diagnose and differentiate asthma from chronic obstructive bronchopulmonary disease (COPD) has been reported (13). However, knowledge of biomarkers to diagnose asthma exacerbation remains uninvestigated. It is necessary to explore diagnostic biomarkers and new therapeutic targets for severe asthma.

In this study, we aimed to explore diagnostic and therapeutic candidate genes associated with severe asthma. We used two machine learning algorithms, least absolute shrinkage and selection operator (LASSO) and support vector machine recursive feature elimination (SVM-RFE), to investigate and validate signature genes for severe asthma based on three public gene expression synthesis (GEO) datasets. We also performed functional enrichment analysis and pathway analysis to identify signaling pathways associated with severe asthma.

2 Materials And Methods

Data Collection and Download
The two asthma gene expression datasets (GSE130499 and GSE63142) analyzed in this study were obtained from the GEO database of the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/geo/). The normal control (NC) and severe asthmatic (SA) (Patients with severe asthma are defined as those patients who require treatment with high-dose inhaled (or systemic) corticosteroids (CS) in combination with a second long-term (controller) medication, includes patients who either maintain control of their disease or who never achieve control) (14) were selected from GSE130499 (38 NC samples and 44 SA samples) and GSE63142 (27 NC samples and 56 SA samples) to identify differential genes. All samples in both datasets were bronchial epithelial cells derived from human.

Data Processing and identification of DEGs

The GSE130499 and GSE63142 datasets were merged and batch variability between them was eliminated using the “SVA” package of R software. Differentially expressed genes were screened with \( P < 0.05 \) and \(|\log \text{FC}| > 0.5\) as screening parameters by using the “limma” package, an R package that processes and analyzes gene expression data (microarray and RNA sequencing) and has become a popular option for finding differential genes.

Gene ontology and pathway enrichment analysis

Gene ontology (GO) and pathway analyses were performed to identify the biological functions of DEGs by using an online database (DAVID Functional Annotation Bioinformatics Microarray Analysis (ncifcrf.gov)). GO analysis investigated the underlying biological processes (BP), cellular components (CC), and molecular functions (MF) of DEGs. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is used to predict the role of protein interaction networks in various cellular activities.

Screening and Validation of Candidate Biomarkers

Two machine-learning algorithms, the support vector machine recursive feature elimination (SVM-RFE) and least absolute shrinkage and selection operator (LASSO) were used in this study to screen candidate biomarkers of asthma. SVM-RFE represents a widely used supervised machine learning protocol for classification and regression, which is performed using the "e1071" package. LASSO regression, a machine learning algorithm with dual characteristics of subset selection and ridge regression, is widely utilized to screen the best variables by finding the lambda value when the classification model error is the least. LASSO regression was performed using the “glmnet” package. The DEGs were screened again using the two machine learning algorithms to find the most relevant candidate biomarkers for asthma. Recipient operating characteristic (ROC) curves were performed to predict candidate biomarkers in the training and validation sets. Finally, the GSE43696 dataset was utilized to validate the differences and of the genes.

3 Results
3.1 Identification of DEGs

To better understand severe asthma, we obtained two microarray datasets, GSE130499 and GSE63142, from the GEO database, then used “SVA” and “limma” packages of R software for data analysis. \( P < 0.05 \) and \(|\log \text{FC}| > 0.5\) were used as screening parameters. A total of 73 DEGs were obtained in this study, which included 45 upregulated genes and 28 downregulated genes. The differential expression gene correlation heatmap is shown in Fig. 1. All DEGs (including logFC and P.Value) are presented in supplementary table 1.

3.2 Functional and Pathway Enrichment Analysis of DEGs

Here, we also performed gene function and pathway analysis of DEGs to conclusively identify the biological significance and enrichment pathways of these genes. GO has revealed that these DEGs are mainly associated with immune and inflammation, such as cellular response to lipopolysaccharide, apoptotic process, defense response to bacterium. In addition, KEGG pathway analysis implies that these DEGs are richly enriched in cytokine-cytokine receptor interaction, staphylococcus aureus infection and Viral protein interaction with cytokine and cytokine receptor. Figure 2 shows the enrichment outcomes of DEGs from GO and KEGG analysis. All items are presented in Table 1.
Table 1
Terms for GO and KEGG analysis; biological processes (BP), cellular components (CC), and molecular functions (MF). The Kyoto Encyclopedia of Genes and Genomes (KEGG)

<table>
<thead>
<tr>
<th>Term</th>
<th>P_Value</th>
<th>Count</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>defense response to bacterium</td>
<td>0.000184847</td>
<td>7</td>
<td>BP</td>
</tr>
<tr>
<td>signal transduction</td>
<td>0.000414725</td>
<td>14</td>
<td>BP</td>
</tr>
<tr>
<td>cellular response to lipopolysaccharide</td>
<td>0.000589616</td>
<td>6</td>
<td>BP</td>
</tr>
<tr>
<td>apoptotic process</td>
<td>0.019768536</td>
<td>7</td>
<td>BP</td>
</tr>
<tr>
<td>positive regulation of transcription from RNA polymerase II promoter</td>
<td>0.025319702</td>
<td>10</td>
<td>BP</td>
</tr>
<tr>
<td>extracellular space</td>
<td>2.47E-11</td>
<td>28</td>
<td>CC</td>
</tr>
<tr>
<td>extracellular region</td>
<td>2.88E-06</td>
<td>22</td>
<td>CC</td>
</tr>
<tr>
<td>secretory granule</td>
<td>0.004438467</td>
<td>4</td>
<td>CC</td>
</tr>
<tr>
<td>extracellular exosome</td>
<td>0.01259661</td>
<td>15</td>
<td>CC</td>
</tr>
<tr>
<td>plasma membrane</td>
<td>0.016153999</td>
<td>27</td>
<td>CC</td>
</tr>
<tr>
<td>growth factor activity</td>
<td>0.000314203</td>
<td>6</td>
<td>MF</td>
</tr>
<tr>
<td>glucocorticoid receptor binding</td>
<td>0.001260677</td>
<td>3</td>
<td>MF</td>
</tr>
<tr>
<td>identical protein binding</td>
<td>0.006778258</td>
<td>14</td>
<td>MF</td>
</tr>
<tr>
<td>chemokine activity</td>
<td>0.013581048</td>
<td>3</td>
<td>MF</td>
</tr>
<tr>
<td>&quot;RNA polymerase II transcription factor activity, ligand-activated sequence-specific DNA binding&quot;</td>
<td>0.015721731</td>
<td>3</td>
<td>MF</td>
</tr>
<tr>
<td>Staphylococcus aureus infection</td>
<td>0.000697845</td>
<td>5</td>
<td>KEGG</td>
</tr>
<tr>
<td>Cytokine-cytokine receptor interaction</td>
<td>0.00138103</td>
<td>7</td>
<td>KEGG</td>
</tr>
<tr>
<td>Viral protein interaction with cytokine and cytokine receptor</td>
<td>0.008666976</td>
<td>4</td>
<td>KEGG</td>
</tr>
<tr>
<td>Estrogen signaling pathway</td>
<td>0.020577876</td>
<td>4</td>
<td>KEGG</td>
</tr>
<tr>
<td>Leishmaniasis</td>
<td>0.042370245</td>
<td>3</td>
<td>KEGG</td>
</tr>
</tbody>
</table>

3.3 Screening of Candidate Biomarkers for severe asthma based on Machine Learning

We used two machine learning algorithms (LASSO regression and SVM-RFE) to screen for potential biomarkers of severe asthma. Total 19 potential biomarkers were identified by using the LASSO regression algorithm (Fig. 3a), while 19 candidate biomarkers were also ascertained using the SVM-RFE
algorithm (Fig. 3b). Afterwards, the intersection of the genes derived from the two algorithms was taken and the final 13 genes were obtained (Fig. 3c). To further confirm the reliability of these biomarkers, we analyzed further dataset GSE67940 from GEO (NC = 20, SA = 88). The results showed that 3 genes (BCL3, S100A14, DDIT4) showed similar trends in expression levels to the previous analysis and were statistically significant (Fig. 4). The differential trends for the other 10 genes are presented in Supplementary Fig. 1.

3.4 Clinical Significance of Candidate Biomarkers for Asthma

Next, we proceeded to further investigation of the diagnostic effectiveness of the 9 genes and verified them with the GSE67940 dataset. As a result, we found that BCL3, DDIT4 and S100A14 had better diagnostic effectiveness, and their AUCs respectively were 0.825, 0.79 and 0.836 (Fig. 5a-c). Furthermore, the AUCs of BCL3, DDIT4 and S100A14 in the GSE67940 dataset were 0.844, 0.793 and 0.797, respectively (Fig. 5d-f). The ROC curves of the other 10 genes are shown in Supplementary Fig. 2.

4 Discussion

Asthma is a complex multifactorial disease with complex and diverse mechanisms including airway inflammation, airway tone control and reactivity (15). Although glucocorticoids currently alleviate asthma, hospitalization or systemic treatment with glucocorticoids is necessary every year owing to the worsening of asthma (16). Lack of valuable biomarkers makes early diagnosis of severe asthma almost impossible. Recently, differentially expressed genes have been considered as candidate pathogenic genes for respiratory diseases, especially asthma. For example, low expression of ITGB4 in airway epithelial cells (AEC) induces asthma airway remodeling (17). However, candidate gene studies on the aberrant expression of relevant differential genes in airway epithelial tissue in severe asthma less investigated.

The aim of this study was to explore potential biomarkers for severe asthma. Two severe asthma gene expression datasets from the GEO database were integrated for comprehensive analysis by using bioinformatics methods. Total 73 differentially expressed genes were screened for further GO function and KEGG pathway analysis. According to the items in GO biological process, the screened genes were mainly enriched in cellular response to lipopolysaccharide, apoptotic process, defense response to bacterium. It is suggested that the genes may have a role in the development of asthma through these biological processes. Asthma is a heterogeneous disease caused by a complex interaction between host genetics, environmental exposures (e.g., allergens), and infectious agents (e.g., viruses and bacteria) (18–20). The role of bacteria in the development and progression of asthma is controversial, however, it may act as a necessary adjunct. Evidence suggested that more than 50% of sputum bacterial cultures are positive performed in severe asthma patients (21). Similarly, a study revealed an increase in intraepithelial neutrophils associated with better lung function in the bronchi of patients with severe asthma (22). The mechanism by which bacteria affect asthma progression may be through
lipopolysaccharide (LPS), also known as endotoxin, an important component of the outer membrane outer leaflet of most Gram-negative bacteria (23). Endotoxin was found to be able to increase the secretion of the Th2 cytokine IL-13 and decrease responsiveness to corticosteroid treatment for asthma. In addition, it has been shown that endotoxin is the leading cause of the shift in asthma phenotype from eosinophilic to neutrophilic by promoting the differentiation of CD4⁺ cells to Th17 cells rather than Th2 cells (24). In terms of asthma treatment, even though some drugs such as CXCR2 inhibitors are not effective in the treatment of severe asthma, showing greatly effective in reducing sputum and blood neutrophils (25). The macrolide, like azithromycin, for asthma significantly reduced the number of asthma attacks and successfully reduced neutrophil-dominant inflammatory biomarkers and exacerbations in severe asthma patients (26), but its bacterial resistance and complication need to be evaluated. Evidently, it is extremely significant to find more targets for asthma treatment. Up to now, research on BCL3 deficiency inducing increased susceptibility of the organism to bacterial infections has been well-reported, among which Streptococcus pneumoniae and Klebsiella pneumoniae, (27, 28) strongly suggesting that BCL3 may have a central role in triggering severe asthma. Apoptosis is another essential mechanism influencing asthma. Studies have confirmed the infiltration of T cells and eosinophils in the bronchial mucosa followed by secretion of pro-inflammatory cytokines TNF-α, IFN-γ, which mediate apoptosis of bronchial epithelial cells and smooth muscle cells (29, 30). Furthermore, the presence of mechanisms of anti-apoptotic mediator release and phagocytosis of apoptotic cells in asthma patients may delay the exacerbation of asthma. One research examined the phagocytosis of apoptotic cells in vitro on bronchoalveolar lavage macrophage from normal subjects, mild-moderate asthma patients and severe asthma patients, finding that macrophage from normal subjects and mild-moderate asthma patients were able to phagocytose apoptotic cells in response to LPS, whereas phagocytosis in severe asthma was defective and detrimental to the regression of inflammation (31). Anti-apoptotic proteins such as protein S have been shown to prevent asthma by shifting the Th1/Th2 balance to Th1 and promoting the secretion of Th1 cytokines (IL-12, TNF-α) from dendritic cells (32). Once the balance between anti-apoptotic, apoptotic cell clearance and pro-apoptotic is disrupted in the body, the symptoms of asthma may be aggravated. The other way to look at this is that it has enlightened to our study since DDIT4 and S100A14 are exactly enriched to this biological process. We may understand the exacerbation process of asthma from the viewpoint of DDIT4 or S100A14 effecting apoptosis.

Our study identified three candidate biomarkers by machine learning methods. Nuclear factor (NF)-κB is a key factor in the normal development and homeostasis of the immune system, controlling the transcription of inflammatory cytokines and chemokines, it is also a protein involved in antigen presentation, and a regulator of cell death and proliferation (33). B-cell lymphoma factor 3 (BCL3) is an atypical member of the ikappa B inhibitor (IkB) family, which activates or inhibits gene transcription by combining with two members of the nuclear factor NF-κB family, p50 or p52 homodimers (34, 35). BCL3 is implicated with the development and progression of many diseases and malignancies (36), such as hematological tumors (37). In addition to this, in inflammatory effects, BCL3 is widely considered as an anti-inflammatory factor that is essential in promoting B cell development, differentiation, survival and proliferation of Th cells, and terminal differentiation of dendritic cell functions (38–40). Mice with
knockout BCL3 have an immunodeficiency in the activation of the NF-κB pathway, lack proper immunity to infection, and have an abnormal inflammatory response (28). Inhibition of NF-κB signaling pathway IκB kinase, phosphorylation of ERK, JNK and P38 MAPK can control IgE and IL-4 production and suppress inflammatory mediators in asthma (41). Deficiency of BCL3 may inhibit the activation of the above pathways thereby exacerbating the onset of asthma. June Guha et al. showed that BCL3 is essential in the initiation or activation of adaptive T-cell immune responses to Toxoplasma gondii by dendritic cells (42). In patients with allergen sensitization, dendritic cells act as specialized antigen-presenting cells that present allergens to T lymphocytes, thereby activating T cell responses to allergens (2). Although the persistent airway inflammation in patients with severe asthma may be caused by an excess of pro-inflammatory molecules in the microenvironment, a similar pathological state may be caused by the absence of counter-regulatory molecules that inhibit the inflammatory response. Therefore, we hypothesize that the role of BCL3 in asthma is mainly to suppress the inflammatory response by activating NF-κB transduction and dendritic cells.

DNA Damage-Induced Transcript 4 (DDIT4), also known as REDD1 or RTP801, is a stress-inducible protein that can be up-regulated at the transcriptional level in response to various stresses (hypoxia, DNA damage, glucocorticoid treatment, etc.) along with its homolog REDD2, both of which negatively regulate the signaling pathway through the mammalian target of rapamycin (mTOR) signaling pathway (43–45). As an mTOR inhibitor, DDIT4 may play a key role in metabolic disorders, neurodegeneration, cancer, aging, and inflammation. Moreover, DDIT4 may play a dual role in immunity and inflammation. There was evidence that the activation, proliferation, and activation of resting T cells depend on the activation of mTOR (46), based on which DDIT4 knockdown can be predicted to promote immune inflammation, however, in the immune cells of diseases such as ulcerative colitis and multiple sclerosis, DDIT4 is overexpressed to facilitate the associated inflammation (47, 48). Besides, among LPS-induced vascular endothelial cell injury and cigarette-stimulated lung injury, DDIT4 negatively regulates signaling pathways including mTOR and NF-κB to induce apoptosis, oxidative stress, and inflammation. Alternatively, DDIT4 may also work without the mTOR pathway. Induced by serum endothelin-1 (ET-1) and hypoxia-inducible factor-1α (HIF-1α), the REDD1 autophagic pathway is activated to lead to enhanced release of neutrophil extracellular traps (NETs), promoting thrombotic inflammation and fibrosis in human systemic lupus erythematosus (SLE) thrombotic inflammation and fibrosis (49). Therefore, we speculate that DDIT4 may play different roles in different types of asthma, and the precise mechanisms demand additional investigation.

S100 calcium-binding protein A14 (S100A14) is a member of the S100 family implicated in many biological processes, which include, for example, regulation of proliferation, differentiation, apoptosis, Ca2+ homeostasis, inflammation, and migration (50). S100A14 is significantly differentially expressed in human diseases, with upregulation in ovarian, pancreatic and breast cancers (51) and downregulation in colorectal tumors and esophageal squamous cell carcinoma (ESCC) (52). Nevertheless, its biological function is currently largely unknown. It was reported that the overexpression of S100A14 promotes the progression of non-small cell lung cancer (53). Agnieszka Pietas et al. showed that S100A14 is
moderately expressed in lung tissue (including normal human bronchial epithelial (HBE) cells), and in addition, they detected an upregulation of the gene at the transcriptional level in lung tumors (54). A report showed that the expression level of S100 proteins (including S100A14) was significantly elevated in NK cells from hiv-exposed seronegative people who inject drugs (HESN-PWID), and in vitro experiments also demonstrated that S100A14 significantly activated NK cells as well as induced tumor necrosis factor-α secretion by monocytes (55). Tumor necrosis factor-α is a pro-inflammatory cytokine that is involved in the pathogenesis of asthma (56). Dong-Fang Meng identified that overexpression of S100A14 could inhibit nasopharyngeal carcinoma cell motility by reversing EMT and inhibiting NF-κB signaling pathway (57). In summary, we speculate that S100A14 may underlie the pathological process of asthma and may help to predict the severity of asthma.

There are several limitations to this study. First, the key differential genes we have identified and their pathways have not been confirmed in asthma by in vitro or in vivo studies; however, this would be an area for further research. Second, whether the differential expression of the three identified genes is associated with the effectiveness of glucocorticoid therapy for severe asthma is currently unknown, and this also needs to be explored in further studies. Finally, although we used only bioinformatics research methods, we applied two different machine learning methods to screen the biomarkers and validated our results using a third data, which strengthens the credibility of our results to some extent.

Declarations

Ethics approval and consent to participate

All data for this study were obtained from public databases. Not applicable.

Consent for publication

Not applicable

Availability of data and materials

The datasets GSE130499, GSE63142 and GSE43696 for this study can be found in the Gene Expression Omnibus(https://www.ncbi.nlm.nih.gov/geo/).

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

Zhangjiajie Yongding District science and Technology innovation plan project (2021111616).

Authors' contributions
Conception or design: ML, FZ, JZ.

Acquisition, analysis, or interpretation of data: FZ, JZ.

Drafting the work or revising: FZ, JZ, ML.

Acknowledgments

This study used the GEO database from National Center for Biotechnology Information (NCBI) as the data source. The authors thank the NCBI and all participants.

References


Figures

Figure 1

Identification of DEGs. (a) Heatmap of DEGs. The red indicates the expression of up-regulation, while the blue indicates the expression of down-regulation. (b) Volcano plot of DEGs.
Figure 2

Bubble plots for GO and KEGG analysis. biological processes (BP), cellular components (CC), and molecular functions (MF).
Figure 3

Screening of Candidate Biomarkers for asthma (a, b) LASSO logistic regression and SVM-RFE algorithm screening diagnostic biomarkers for asthma. (c) The Venn diagram showing the intersection of the candidate biomarkers screened by the two algorithms.
Figure 4

Validation of candidate biomarkers expressions in the GSE43696 dataset

Figure 5

Diagnostic significance of candidate biomarkers in severe asthma. (a-c) ROC curves of BCL3, DDIT4 and S100A14 in the combined dataset of GSE130499 and GSE63142, (d-f) ROC curves of BCL3, DDIT4 and
S100A14 in the GSE43696 dataset.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf