Original Article

Two gene set variation indexes as potential diagnostic tool for sepsis

Junyu Lu1*, Qian Li2*, Zimeng Wu2, Zhimei Zhong2, Pan Ji2, Hongyuan Li2, Cuiling He2, Jihua Feng2, Jianfeng Zhang2

1Intensive Care Unit, The Second Affiliated Hospital of Guangxi Medical University, Nanning 530007, China; 2Department of Emergency Medicine, The Second Affiliated Hospital of Guangxi Medical University, Nanning 530007, China. *Equal contributors.

Received January 31, 2020; Accepted May 26, 2020; Epub June 15, 2020; Published June 30, 2020

Abstract: Accurate diagnosis of sepsis remains challenging, new markers or combinations of markers are urgently needed. In the present study, we screened differentially expressed genes (DEGs) between sepsis and non-sepsis blood samples across three previously published gene expression data sets. Common upregulated and downregulated DEGs were ranked according to their average functional similarity. The ten genes (OLFM4, ORM1, CEP55, S100A12, S100P, LRG1, CEACAM8, MS4A4A, PLSCR1, and IL1R2) with the largest average functional similarity among the common upregulated genes and another ten genes (THEMIS, IL2RB, CD2, IL7R, CD3E, KLRB1, PVRIG, CCRR3, TGFBR3, and PLEKHA1) with the largest average functional similarity among the common downregulated genes were separately identified as the upregulated crucial gene set and the downregulated crucial gene set. Gene set variation analysis (GSVA) was used to obtain the GSVA index of each sample against the two crucial gene sets. Both the two crucial GSVA indexes may be robust markers for sepsis with high area under ROC curve. The diagnostic utility of the upregulated GSVA index was validated in another independent data set. Functional analyses revealed several sepsis-related pathways. In conclusion, we proposed two sepsis-related gene sets across multiple data sets and created two GSVA indexes with promising diagnostic value.

Keywords: Sepsis, systemic inflammatory response syndrome, biomarker, gene set variation analysis

Introduction

Sepsis is a heterogeneous disease involving a dysregulated systemic response to infections caused by bacteria, fungi or viruses, which may lead to organ dysfunction [1]. No specific treatment currently exists for sepsis [2]. Sepsis is a common illness and one of the ten leading causes of death worldwide [3]. Even though there has been some progress in the treatment of sepsis in the last years, the mortality rate of sepsis is still terrible, the risk of death from sepsis is as high as 30%, and risk of death rises to 50% in the case of severe sepsis and around 80% in the case of septic shock [4]. The diagnosis of sepsis can be complicated due to its clinical similarities with systemic inflammatory response syndrome (SIRS), which can also lead to multiple organ failure and death [5]. SIRS can be caused by infection, trauma, burns, pancreatitis, or a variety of other injuries [6]. A previous study proposed that screening ward patients using SIRS criteria for identifying those with sepsis would be impractical [7], thus, one of the challenges facing sepsis research is to determine which patients are truly infected.

Antibiotics are currently the standard treatment for sepsis [8]. However, delays in the use of antibiotics can result in increased mortality [9]. Since sepsis and non-bacterial SIRS are hard to distinguish in the early stages, their clinical symptoms were very similar. Antibiotics are often abused, which may further contribute to the spread of drug resistance. Blood purification techniques, immunomodulatory drugs and treatments targeting other systems including the heart, endothelial cells or coagulation cascades are also used for sepsis management [10], but they lack specificity. The biomarkers such as c-reactive protein (CRP) [11] and procalcitonin (PCT) [12] could be used to distinguish bacterial sepsis from other inflammatory conditions, but they had some limitation [13]. There was another biomarker, lactate which
Two gene set variation index as biomarker of sepsis

Could use to diagnosis sepsis, however, it lacks specificity [14]. Therefore, in order to establish a targeted treatment for sepsis, it is necessary to first establish a diagnostic model able to distinguish sepsis and non-sepsis conditions.

In this study, we identified differentially expressed genes (DEGs) between sepsis and non-sepsis blood samples across multiple data sets and created two gene sets separately comprised of ten genes as diagnostic models of sepsis. The receiver operating characteristic (ROC) curve analysis was used to explore the potential diagnostic value of these two gene set variation analysis indexes for sepsis, and further validated this in an independent data set. Moreover, we explored the potential role of these genes in sepsis pathways using the Kyoto Encyclopedia of Genes and Genomes (KEGG), and their potential involvement in dysregulated systemic responses.

Materials and methods

Data collection and processing

Gene Expression Omnibus (GEO) datasets were downloaded from https://www.ncbi.nlm.nih.gov/geo/ and used to compare gene expression in sepsis and non-sepsis. GSE57065 [15] included whole blood gene expression profiles of 28 sepsis samples and 25 non-sepsis healthy samples, based on the GPL570 platform. Whole blood gene expression profiles of GSE69528, which included 53 sepsis samples and 85 non-sepsis (uninfected type 2 diabetes mellitus, uninfected healthy and septicemic melioidosis), were based on the GPL-10558 platform. Whole blood gene expression profiles of GSE5923 [9], based on GPL-16791, included 51 sepsis samples and 22 non-sepsis (healthy) samples. The normalize Between Arrays function in the limma package in R [16] was used to normalize the gene expression expression profiles. If a gene corresponded to multiple probes, the average expression value of these probes was considered to reflect the expression of the gene. The workflow of the present study was shown in Figure 1.

**Figure 1.** Flowchart of the present study.

**Principal component analysis (PCA) and differentially expressed gene (DEG) analysis**

To evaluate whether differences in gene expression patterns in whole blood can distinguish between sepsis and non-sepsis, the princomp [17] function was used to perform PCA [18], and the ggbiplot package in R [19] was applied to visualize the results. Compared with non-sepsis, DEG were screened using limma package, and |log fold change (FC)| > 1.5 and P adjusted by the false discovery rate (FDR) < 0.01 were considered significant. We applied Venn diagram analysis [20] to find upregulated and downregulated genes common to the three data sets.

**Identification of crucial genes, Gene Set Variation Analysis (GSVA) and ROC curve analysis**

Crucial genes were screened by semantic similarity in the upregulated and downregulated gene sets of all three GEO data sets. According to the semantic similarities of Gene Ontology (GO) terms used for gene annotation, we ranked the gene as used to inside the interactome by
the average functional similarities between the gene and its interaction partners. Genes with the highest average functional similarity were considered crucial genes [21]. GSVA was used to assess the relative enrichment of gene sets across samples using a non-parametric approach [22]. The function of gsva was used with gsva method to score each sample for the crucial gene sets, and then each sample received a value of the upregulated GSVA index and a value of the downregulated GSVA index. In addition, ROC curve analysis was performed using pROC package [23] to evaluate the diagnostic value of the GSVA indexes for sepsis in GSE57065, GSE69528 and GSE95233 datasets.

Validation of diagnostic utility of the upregulated GSVA index

GSE123729, which comprised 15 sepsis samples and 27 non-sepsis samples (11 presurgical and 16 SIRS) based on GPL21970, was used as the validation set. Similarly, each sample was assigned a value for the upregulated GSVA index using the GSVA method. The ROC curve analysis was also applied to evaluate the diagnostic value of the upregulated GSVA index for sepsis in GSE123729. PCA was used to explore the utility of the classification between sepsis and non-sepsis using the crucial upregulated genes.

Functional enrichment analysis

Standard pathway analysis may have limited ability to reveal regulatory mechanisms of key genes hidden in long pathways or sub-pathways [24]. Therefore, in order to identify potential risk sub-pathways in sepsis, sub-pathway enrichment analysis was performed using the Sub-pathway Miner [25] package in R for the upregulated and downregulated genes in all three data sets. Gene set enrichment analysis (GSEA) [26] was performed using the normalized gene expression profile to explore KEGG pathways related to sepsis. GSEA software was used for this analysis (http://software.broadinstitute.org/gsea/index.jsp), and c2.cp.kegg.v6.2.symbols.gmt, which come from the Molecular Signatures Database (MSigDB) was used as the reference gene set [26, 27]. A nominal value of \( P < 0.05 \) was considered statistically significant. The ggplot2 package [28] in R was used to visualize the results of the GSEA.

Results

DEGs in sepsis

The results of the PCA indicated that sepsis and non-sepsis whole blood samples had significantly different gene expression patterns (Figure 2A). Compared with non-sepsis samples, a total of 298 DEGs were found in GSE57065, 174 of which were upregulated and 124 downregulated. A total of 343 DEGs were observed in GSE69528, 198 upregulated and 145 downregulated. A total of 428 DEGs were found in GSE95233, 257 regulated and 171 downregulated (Figure 2B).

A total of 88 DEGs were upregulated in all three data sets, while a total of 32 downregulated DEGs were common to the three data sets (Figure 2C). The heatmap showing upregulated and downregulated gene sets shows their potential to distinguish between sepsis and non-sepsis (Figure 2D).

Crucial GSVA index may have diagnostic for sepsis

We ranked genes by their average functional similarity relationships with other genes within the interactome. In the three GEO data sets, the ten genes with the largest average functional similarity among the common upregulated genes of sepsis were OLFM4, ORM1, CEP55, S100A12, S100P, LRG1, CEACAM8, MS4A4A, PLSCR1 and IL1R2 (Figure 3A), while THEMIS, IL2RB, CD2, IL7R, CD3E, KLRB1, PVRIG, CCR3, TGFBR3 and PLEKHA1 showed the largest functional similarity among the common downregulated genes (Figure 3B). The heatmap showed that these DEG expression patterns could distinguish sepsis from non-sepsis (Figure 3C). In the ROC curve analysis, the GSVA index of upregulated genes had an AUC = 0.9849 in GSE57065, an AUC = 0.8276 in GSE69528, and an AUC = 0.7669 in GSE95233 (Figure 3D, 3E). The GSVA index of downregulated genes had an AUC = 0.9712 in GSE57065, AUC = 0.8469 in GSE69528, and AUC = 0.6087 in GSE95233.

Validation of upregulated GSVA index in an independent data set

The ROC analysis suggested that the upregulated GSVA index may be a robust marker for...
Two gene set variation index as biomarker of sepsis

A

B

C

D

GSE57065

GSE69528

GSE95233

GSE57065

GSE69528

GSE95233

GSE57065

GSE69528

GSE95233

GSE57065

GSE69528

GSE95233
Two gene set variation index as biomarker of sepsis

Figure 2. Differentially expressed genes in sepsis compared to non-sepsis. A. Principal component analysis of sepsis vs. non-sepsis. B. Manhattan plot of differentially expressed genes in sepsis and non-sepsis. The top ten upregulated and downregulated genes with the highest significance (ranked by P value) are highlighted. Gray represents genes that are not significantly differentially expressed. C. Venn diagram showing upregulated genes and downregulated genes. D. Expression heatmap of upregulated and downregulated genes common to the three data sets.

Figure 3. The two crucial gene set variation analysis (GSVA) indexes in sepsis. A. Crucial upregulated genes. B. Crucial downregulated genes. C. Heatmap of crucial genes and GSVA index in all three data sets. D. Receiver operating characteristic (ROC) curve analysis of upregulated GSVA index. E. ROC curve analysis of downregulated GSVA index.

sepsis, with AUC = 0.929 (Figure 4A) in the independent data set. Compared with non-sepsis samples, the upregulated GSVA index in sepsis samples was significantly higher (Figure 4B). Subsequently, we found that sepsis and non-sepsis could be well differentiated in GSE123729 based on the expression patterns of the crucial upregulated genes (Figure 4C).

Dysfunctional pathways in sepsis

The sub-pathway enrichment analysis indicated that the common upregulated genes were involved in various sepsis-related pathways, such as coagulation cascades, P53 signaling pathways and MAPK signaling pathways. The common downregulated genes were involved in
Two gene set variation index as biomarker of sepsis

Jak-STAT, PI3K-Akt and T cell receptor signaling pathways (Figure 5A). The GSEA results also confirmed that complement and coagulation cascades and P53 signaling pathway were enriched in sepsis samples in all three data sets. In contrast, allograft rejection, autoimmune thyroid disease and T-cell receptor signaling pathway were enriched in the non-sepsis samples in the three data sets (Figure 5B).

Discussion

In original study of GSE57065, the correlation between gene expression pattern and clinical severity of sepsis was evaluated by severity score [29]. In original study based on GSE-95233, several candidate biomarkers, including CX3CR1 and Lirb 2, were identified as prognostic biomarkers of sepsis [9]. The aim of original study based on GSE69528 was to find biomarkers to distinguish bacterial-induced sepsis and sepsis caused by other pathogens [30]. Compared with these three original studies, we pay more attention to the variation of the crucial gene sets rather than the aberration of a single molecule. In the present study, we screened common upregulated or downregulated genes across three gene expression data sets of patients with sepsis in comparison with non-sepsis individuals. We identified a crucial gene set according to the average functional similarity among genes upregulated or downregulated across the various data sets. The crucial upregulated gene set comprised OLFM4, ORM1, CEP55, S100A12, S100P, LRG1, CEACAM8, MS4A3A, PLSCR1 and IL1R2. Previous studies have shown that OLFM4 can negatively regulate the defense response against bacterial infections [31]. ORM1 was found to be significantly upregulated in sepsis [32, 33], similarly to S100P [34] and LRG1 [35], and they may have a role as sepsis biomarkers. S100A12 and CEACAM8 may be involved in natural immunity against sepsis [36, 37]. PLSCR1 has been shown to take part in innate protective mechanisms against a bacterial pore-forming toxin [38]. IL1R2 is a potential biomarker for diagnosis of sepsis [39]. In our present study, CEP55 and MS4A4A were upregulated in sepses and identified as crucial genes, this indicates CEP55 and MS4A4A may be closely associated with sepsis. OLFM4 could promote leukocyte mediated migration, neutrophil activation and degranulation process [40]. CEP55 can cause the proliferation of cytotoxic T lymphocytes in vivo [41]. S100A12 plays an important role in promoting the formation of osteoclasts, which have a certain effect on systemic inflammation [42]. High expression of PLSCR1 could inhibit phagocytosis of macrophages [43]. Tumor infiltrating regulatory T cells (Treg) can inhibit tumor antigen-specific T cells, and IL1R2 was found to be expressed on the cell surface of Treg [44]. Most of the genes are found to be related to immune cells, Therefore, the variation of the crucial gene sets may reflect the response to sepsis in the host immune system.

The crucial downregulated gene set included THEMIS, IL2RB, CD2, IL7R, CD3E, KLRB1, PVRIG, CCRR3, TGFBR3 and PLEKHA1. Previous work found that IL2RB and CD3E were negatively correlated with sepsis organ failure and mortality [45], which is consistent with our study. In other study [46], IL7R expression was also found to be downregulated in sepsis. TGFBR3 was suggested to help identify a “bacteremia-prone” phenotype in sickle cell anemia [47]. There were few studies reporting a relationship of THEMIS, CD2, KLRB1, PVRIG, CCRR3 or PLEKHA1 with sepsis. Our result suggests that the downregulation of these genes

Figure 4. Validation of the upregulated gene set variation analysis (GSVA) index using the GSE123729 expression data set. A. ROC curve analysis of the upregulated gene set. B. Comparison of GSVA score in sepsis vs. non-sepsis in GSE123729. C. Principal component analysis of the crucial upregulated gene set in GSE123729.
Two gene set variation index as biomarker of sepsis

Figure 5. Functional enrichment analysis. A. Sub-pathway analysis for common upregulated/downregulated genes in sepsis. B. Pathways enriched in sepsis or non-sepsis in the three gene expression data sets.

may be response of the host to sepsis. THEMIS was the key in the development of T-cell [48]. IL12RB was associated with impaired T cell [49]. CD2 is a glycoprotein on the surface of T lymphocytes, and its abnormality may be due to the activation of lymphocytes [50]. IL7R could develop into plasma like dendritic cells [51]. CD3 affinity is closely related to the distribution of T cell rich tissue [52]. PVRIG was a checkpoint receptor that can induce T cell enhancement [53]. TGFB3 was found to be preferentially expressed in CD4+ T cells [54]. Thus, the aberrant expression of genes may be caused by the blood cell composition change.
Two gene set variation index as biomarker of sepsis

Nevertheless, compared with the cell population level, our present study may access these subtle changes due to the variation of crucial gene sets were determined. In clinical practice, procalcitonin (PCT) assays are widely used for the diagnosis of sepsis, [55] although a previous meta-analysis reported that PCT has a sensitivity of only 76% and a specificity of 70% for bacteremia [56]. Therefore, more specific markers are urgently needed for effective diagnosis of sepsis. In the present study, we created two GSVA indexes and found that they may be robust biomarkers for sepsis, as suggested by their high AUC. The upregulated GSVA was also validated with an independent data set. The two GSVA indexes may be worthy of further exploration.

The occurrence of certain diseases is not always caused by the abnormality of the whole pathway involved in the biological process, but it can be caused by the dysfunction of a sub-pathway [57]. Several upregulated genes in sepsis were enriched in the sub-pathway analysis, including complement and coagulation cascades, which is consistent with other research [58]. Complement and coagulation cascades can facilitate the containment and destruction of pathogens to protect against bacterial spreading within the body. Inhibition of p53 expression can significantly inhibit cardiomyocyte apoptosis induced by sepsis [59], and p53 signaling was upregulated in our sepsis data, which means that it may promote the development of the disease. Kidney transplant recipients developing sepsis showed inferior patient survival and allograft function, and therefore the identification of differences in alloreactivity may be useful to identify transplant recipients at increased risk [60]. Chronic critical illness from sepsis has been associated with an enhanced T-cell receptor response [61]. These pathways may play a role associated with sepsis according to our analyses.

As our best knowledge, our study is the first study to explore the diagnostic value of gene set variation index in sepsis. However, our work has several limitations. Firstly, we were unable to validate the downregulated gene set using another independent data sets. One obvious reason is the heterogeneity of the three non-septic samples which included healthy volunteers, patients with type 2 diabetes mellitus. Secondly, the upregulated and downregulated GSVA indexes have different AUC values in different data sets due to the different numbers of sample in the data sets. However, all the values of AUC were above 0.7 for the upregulated GSVA index in the all data sets. In addition, we did not experimentally validate the potential of the two gene sets for the diagnosis of sepsis. Therefore, it is not clear whether these genes are causal or merely markers for sepsis. Nevertheless, our study may provide a preliminary basis for the exploration of new biomarkers for the diagnosis of sepsis.

In conclusion, we identified two crucial gene sets across multiple data sets in sepsis patients, and we created two GSVA indexes with promising diagnostic value for sepsis.

Acknowledgements

The authors would like to thank Life-Ontology Biological Technology (Nanning, China) for assisting with bioinformatics analysis. This study was supported by the National Natural Science Foundation of China (Grant numbers: 81660132 and 81960343); the Key Research and Development Project of Guangxi (Grant number: Guike AB17195002); Guangxi Natural Science Foundation (Grant numbers: 2017GXNSFBA198043 and 2017GXNSFAA198249); the Scientific Research Funding from Population and Family Planning Commission of Guangxi Zhuang Autonomous Region (Grant number: S2017009); the Reserve Cadre Training Program Science Foundation of the Second Affiliated Hospital of Guangxi Medical University (Grant number: HBRC201805) and the High-level Medical Expert Training Program of Guangxi “139” Plan Funding (Grant number: G201903027).

Disclosure of conflict of interest

None.

Address correspondence to: Drs. Jihua Feng and Jianfeng Zhang, Department of Emergency Medicine, The Second Affiliated Hospital of Guangxi Medical University, No 166 Daxuedong Road, Nanning 530007, Guangxi, China. Tel: +86-7713277-161; Fax: +86-7713277285; E-mail: fengjihua2013@sina.com (JHF); zhangjianfeng@stu.gxmu.edu.cn (JFZ)
Two gene set variation index as biomarker of sepsis

References

Two gene set variation index as biomarker of sepsis


Two gene set variation index as biomarker of sepsis


