Computers in Biology and Medicine 64 (2015) 24–29

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Prediction of feature genes in trauma patients with the TNF rs1800629 A allele using support vector machine

Guoting Chen [1], Ning Han [1], Guofeng Li, Xin Li, Guang Li, Yangzhou Liu, Wei Wu, Yong Wang, Yanxi Chen, Guixin Sun, Zengchun Li, Qinchuan Li *

Department of Emergency Surgery, East Hospital, Tongji University School of Medicine, No. 150, Jimo Road, Shanghai 200120, China

ARTICLE INFO

Article history:
Received 1 February 2015
Accepted 1 June 2015

Keywords:
Severe trauma
Tumor necrosis factor-α rs1800629 A allele
Support vector machine
Feature gene
Classification

ABSTRACT

Background: Tumor necrosis factor (TNF)-α variant is closely linked to sepsis syndrome and mortality after severe trauma. We aimed to identify feature genes associated with the TNF rs1800629 A allele in trauma patients and help to direct them toward alternative successful treatment.

Methods: In this study, we used 58 sets of gene expression data from Gene Expression Omnibus to predict the feature genes associated with the TNF rs1800629 A allele in trauma patients. We applied support vector machine (SVM) classifier model for classification prediction combining with leave-one-out cross validation method. Functional annotation of feature genes was carried out to study the biological function using database for annotation, visualization, and integrated discovery (DAVID).

Results: A total of 133 feature genes were screened out and was well differentiated in the training set (14 patients with variant, 15 with wild type). Moreover, SVM classifier peaked in predictive accuracy with 100% correct rate in training set and 86.2% in testing set. Interestingly, functional annotation showed that feature genes, such as *HMOX1* (heme oxygenase (decycling) 1) and *RPS7* (ribosomal protein S7) were mainly enriched in terms of cell proliferation and ribosome.

Conclusion: *HMOX1* and *RPS7* may be key feature genes associated with the TNF rs1800629 A allele and may play a crucial role in the inflammatory response in trauma patients. Moreover, the cell proliferation and ribosome pathway may contribute to the progression of severe trauma.

1. Introduction

Severe trauma remains to be a global health problem and the second leading cause of death under the age of 45 years [1]. Patients subjected to severe trauma are at a risk of developing sepsis and subsequent organ dysfunction [2]. Moreover, severe trauma and sepsis are related to a high rate of immune irregularities due to an acute inflammatory response in injured sites after encountering trauma or sepsis [2]. The inflammatory response is the host's response to extraneous threat, such as invading pathogen or tissue damage, and subsequent up-regulation of pro-inflammatory cytokines, chemokines and immune cells into tissues [3,4].

The cytokine response plays an important role in the development of severe trauma in that pro-inflammatory cytokines including tumor necrosis factor α (TNF-α), interleukin (IL)-6, 8, 12 and granulocyte-macrophage colony-stimulating factor (GM-CSF) have been shown to release excessively. TNF-α is considered as an essential component in immune response to inflammation that primarily produced by immune cells including macrophages, monocytes and other non-immune cells [5,6]. Currently, TNF-α is shown to act as a key intermediary in inflammatory response that mediates a cascade of cytokines. Furthermore, TNF-α is responsible for organ failure as well as increased risk of sepsis after multiple injuries [7,8]. A previous study has revealed that elevated circulating levels of TNF-α are observed in response to severe trauma [9]. Besides, several single nucleotide polymorphisms (SNPs) within the promoter region or coding sequence of TNF-α have been thought to increase the risk for severe sepsis after trauma and may be associated with outcome of trauma [10,11]. Menges et al. also indicated that variations of the gene encoding for TNF-α (TNF-α variants) were related to sepsis syndrome and death following severe trauma [12]. Therefore, TNF-α variants may be key factors during the development of severe trauma, and identification of feature genes related to TNF-α variants may help improve therapy of trauma. However, classification of trauma patients based on TNF-α variants has never been reported in

* Corresponding author. Tel./fax: +86 21 38804518.
E-mail address: qinchuan_li@163.com (Q. Li).
[1] Guoting Chen and Ning Han contributed equally in this work.

http://dx.doi.org/10.1016/j.compbiomed.2015.06.002
0010-4825/

previous study, as well as feature genes associated with TNF-α variants.

Various feature gene identification methods in handling microarray data have been applied, especially support vector machine (SVM) which is one of the most effective method in classification [13]. Previously, a prospective cohort study based on microarray data GSE5760 was performed to confirm that TNF-α variants were associated with increased occurrence of sepsis syndrome in trauma patients and the peripheral blood transcriptome in patients with or without the TNF rs1800629 A allele was different [12]. Moreover, microarray data GSE5760 was also used to investigate candidate genes and miRNAs for elucidate the effect of TNF rs1800629 A allele on sepsis syndrome in multiple trauma patients [14]. In contrast to this, we used microarray data GSE5760 in this study and predicted feature genes associated with the TNF rs1800629 A allele by applying the SVM method in conjunction with leave one out cross validation (LOOCV). Moreover, functional annotation of feature genes was conducted. The objective of our study was to predict feature genes associated with TNF rs1800629 A allele in trauma patients. Our study will help to elucidate the molecular mechanism of severe trauma and is crucial for drug target selection in clinical application.

2. Materials and methods

2.1. Data source

The dataset GSE5760 deposited by Menges et al. [12] was downloaded from GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5760) based on the platform of GE Healthcare/Amersham Biosciences CodeLink UniSet Human I Bioarray. A total of 28 patients were used for array analysis, including 12 patients with the TNF rs1800629 A allele (MUT1-12) and 16 patients without the TNF rs1800629 A allele (WT1-16). The transcpriptome from peripheral blood from 28 patients was firstly examined using the CodeLink UniSet Human 10 K Bioarrays (GE Healthcare, Freiburg, Germany). Only whole blood samples were used for this investigation upon admission to the intensive care unit (ICU). Each patient sample was hybridized in duplicate on microarrays (label-extract technical replicate). The arrays were designated WTx.y or MUTx.y with WT for patients without and MUT for patients with the TNF rs1800629 A allele, with *x* for the patient-ID (1, 2, 3 and 4) and *y* for the technical replicate number (1, 2 or 3). Microarray results were validated by quantification of mRNA using TaqMan® technology. After microarray quality analysis, a set of 58 arrays containing 28 blood samples with the TNF rs1800629 A allele and 30 samples without the TNF rs1800629 A allele were subjected to microarray analysis. In this study, the first 29 out of 58 samples were used as training set (14 samples with rs1800629 A allele and 15 samples without variant based on number order of GSM) and the remaining samples were used as testing set. A schematic diagram of analysis in this study was shown in Fig. 1.

2.2. Data preprocessing

Raw expression data were quantile normalized using a Robust Multi-array Averaging (RMA) [15] method. The probe data were converted into gene symbol based on probe annotation file. If multiple probes corresponded to the same gene symbol, the mean value was calculated as the gene expression value of this gene. Data were then log 2 transformed [16] for further analysis.

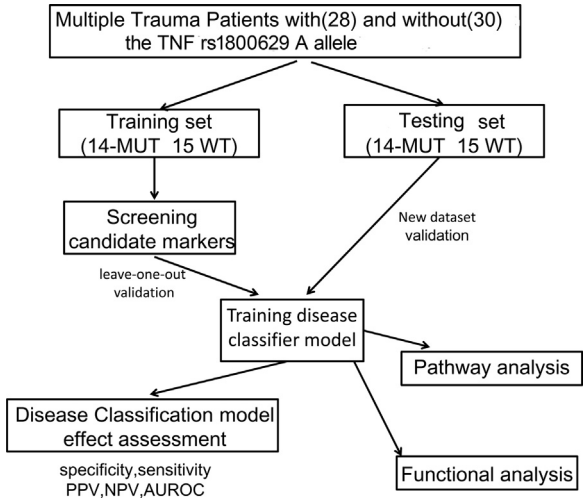

Fig. 1. Schematic diagram of analysis for prediction feature genes in trauma patients with tumor necrosis factor-α rs1800629 A allele.

2.3. Identification of feature genes in training set

To distinguish samples with or without TNF-α rs1800629 A allele, Limma package [17] was applied to identify feature genes with differentially expressed in training set. *P* values were adjusted for multiple testing by Benjamini–Hochberg procedure as the false discovery rate (FDR) [18]. Genes with FDR less than 0.05 and fold change (FC, mutation/wild type) more than 1.5 were considered as significant.

2.4. Classification prediction analysis

SVM is employed as the predictor, which has proven to be a powerful machine learning technique, especially for classification [19,20]. In this study, we performed our classification analyses using a SVM classifier (kernel: radial basis, gamma=0.0075, *C* parameter default value=1) based on feature genes associated with the TNF rs1800629 A allele. SVM classifier was implemented by the e1071 package in R (http://cran.r-project.org/package=e1071) [21]. Moreover, LOOCV [22,23] which involves iteratively leaving out one sample as testing sample until no remaining training samples was applied to estimate the generalized performance of our classifier. Thus, for *N* samples, the LOOCV would train the classifier *N* times.

2.5. Evaluation metrics for SVM

In order to accurately evaluate the classification prediction of SVM classifier, based on the results of LOOCV assessment, the classification models for samples with or without TNF-α variant were comprehensively evaluated based on the calculation of sensitivity, specificity [24], positive predictive value (PPV), negative predictive value (NPV) [25] and the area under the receiver operating characteristic (AUC) curve [26]. If the five parameters equal 1, a perfect classifier can be found. However, if they are equal to 0.5, the classifier has no discriminative power at all.

2.6. Functional annotation of the expression data

To further study the functions of feature genes, Gene Ontology (GO) functional analysis was conducted using database for annotation, visualization, and integrated discovery (DAVID, available at http://david.abcc.ncifcrf.gov/) [27] online tool. Kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment analysis was