Artificial Intelligence in Medicine 76 (2017) 1–6

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Protein coding gene *CRNKL1* as a potential prognostic biomarker in esophageal adenocarcinoma

Zhen Li [a], Qianlan Yao [a], Songjian Zhao [a], Zhen Wang [b,*], Yixue Li [a,b,*]

[a] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai, China
[b] Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

ARTICLE INFO

Article history:
Received 12 December 2016
Received in revised form 12 January 2017
Accepted 19 January 2017

Keywords:
Esophageal adenocarcinoma (EAC)
Prognosis
Biomarker
CRNKL1

ABSTRACT

Background: Esophageal adenocarcinoma (EAC) is one of the most aggressive gastroesophageal cancers. *PTGS2*, *EGFR*, *ERBB2* and *TP53* are the traditional EAC prognostic biomarkers, but they are still limited in their ability to effectively predict the overall survival.
Objectives: To identify an improved biomarker for predicting the prognosis of EAC by using the expression profile.
Materials and methods: Differential co-expression analysis and differential expression analysis were performed to identify the related genes of EAC. The 5-fold cross-validation was used to select a prognostic biomarker from the 532 EAC related genes.
Results: CRNKL1 was identified as a prognostic biomarker to predict the survival of EAC patients. It could significantly stratify EAC patients into high-risk and low-risk groups and was much better than the traditional biomarkers. Furthermore, ROC curve also verified that CRNKL1 with the highest area under the curve (AUC), reaching a sensitivity of 83.33% and a specificity of 78.57%.
Conclusions: Our research proposed that CRNKL1 might be a novel prognostic biomarker with better predictive ability by comparing with the traditional biomarkers, which provided a preferable opportunity in the clinical applications of EAC.

1. Introduction

Esophageal adenocarcinoma (EAC) is a subtype of esophageal cancer that is associated with overweight, obesity, and chronic gastroesophageal reflux disease [1–3]. In recent decades, EAC increases faster than other malignancies in western countries [4], and only 16% of patients could survive more than 5 years after diagnosis and the median survival time is less than 1 year [5]. Despite of recent advances in prognosis and therapies, the survival rates and time were not largely improved. EAC is usually detected in the middle-late stage, leading to poor survival [6,7]. Because the pathogenesis of EAC is concealed, it generally miss the best treatment opportunity. If it can be caught in time, it will be apparently treated with relative ease. Therefore, it is necessary to develop effective biomarkers for predicting the prognosis of EAC patients, which is able to give a survival benefit.

A biological marker can afford an indication of the disease condition, whether normal or abnormal, which should be sensitive, specific and cost-effective. In current clinical practice, *PTGS2*, *EGFR*, *ERBB2* and *TP53* were considered as the optimal prognostic EAC biomarkers [7–11]. COX-2, an induced enzyme, is encoded by the *PTGS2* gene, and its expression affects survival outcomes of patients in EAC [12,13]. *EGFR* and *ERBB2* are members of the *EGFR* family, and involved in at least 3 different oncogenic pathways. Increased *EGFR* expression is correlated with higher tumor stages and worse overall survival [10,14]. HER-2 is encoded by the *ERBB2* gene, its overexpression and amplification were all associated with poor cancer-specific survival [15,16]. *TP53* is a typical tumor suppressor gene and its overexpression is commonly observed in adenocarcinoma of the esophagus, but its prognostic value appears limited [17,18]. However, these biomarkers can only improve the current predictive efficiency of EAC by a low specificity and a restricted sensitivity.

The aim of this study was to investigate the expression pattern of EAC genes and find an improved biomarker for prognosis bet-

* Corresponding authors at: Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
E-mail addresses: zwang01@sibs.ac.cn (Z. Wang), yxli@sibs.ac.cn (Y. Li).

http://dx.doi.org/10.1016/j.artmed.2017.01.002
0933-3657/

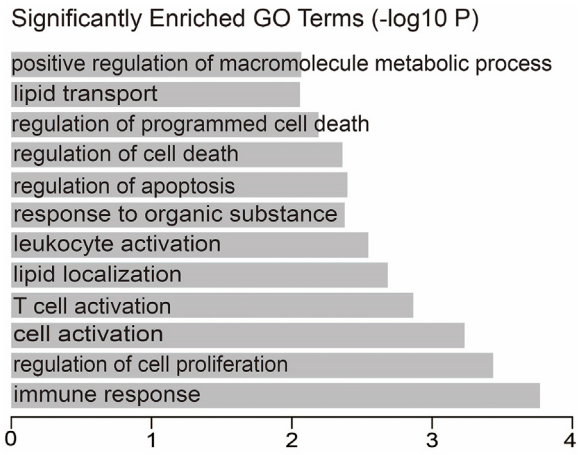

Fig. 1. The biological functions of DCGs. The cutoff of functional enrichment analysis for DCGs was Benjamin adjust *p*-value < 0.01.

ter than the well-studied biomarkers. Differential co-expression analysis and differential expression analysis were used to identify the EAC related genes. The 5-fold cross-validation was performed to select a potential prognostic biomarker. Both in the training and testing datasets, the prognostic biomarker could divide EAC patients into two groups with markedly different outcomes. Compared with traditional EAC biomarkers, the new one achieved higher sensitivity and specificity. All findings suggested that our study efficiently identified a prognostic marker with a better prognostic power in EAC.

2. Materials and methods

2.1. Dataset

The normalized gene expression dataset of EAC was downloaded from UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu/proj/site/hgHeatmap/) [19]. The cancer datasets of UCSC Cancer Genomics Browse were derived from The Cancer Genome Atlas (TCGA) database. The EAC dataset totally included 87 EAC and 10 normal tissue samples, which were involved in 19063 genes (mRNAs). After removing the genes with more than 80% missing values in the 97 samples, 16458 genes were remained for analysis. We randomly selected 55 EAC samples as the training dataset, and the remaining 32 EAC samples as testing dataset.

2.2. EAC related genes

Differential co-expression analysis for the expression profile of the training dataset was conducted in R environment (V3.2.3) using

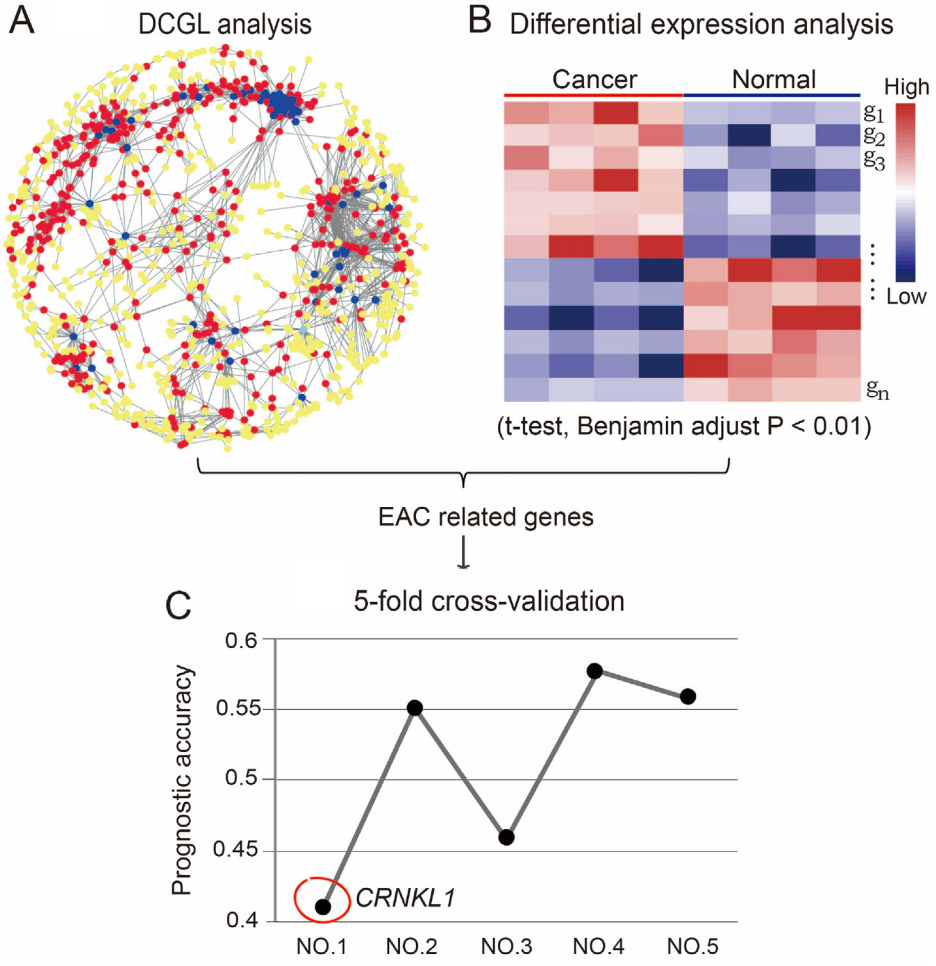

Fig. 2. An integrative pipeline for identification of prognostic biomarker. (**A**) Part of differential co-expression network. The red ellipses were DCGs, the blue ellipses were the important DCGs that regarded as the candidate EAC related genes, and the yellow ellipses were the rest genes. (**B**) Differential expression analysis was performed to identify differential expressed genes from the important DCGs (Benjamin adjust p-value < 0.01) by using the expression profile. (D) The 5-fold cross-validation was used to identify a biomarker for predicting the survival of EAC patients. HR analysis was used to calculate the prognostic accuracies of candidates. The gene with the highest prognostic accuracy was shown in the plot.