Journal of Bone Oncology 7 (2017) 18–22

Contents lists available at ScienceDirect

Journal of Bone Oncology

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

Research Paper

Prediction of optimal gene functions for osteosarcoma using gene ontology and microarray profiles

Xinrang Chen

Pediatric Surgery, First Affiliated Hospital of Zhengzhou University, NO. 1 Jianshe East Road, Zhengzhou, Henan 450052, People's Republic of China

ARTICLE INFO

Keywords:
Osteosarcoma
Spearman correlation coefficient
Gene ontology
Area under the curve
Differential co-expression network
Guilt by association

ABSTRACT

In the current study, we planned to predict the optimal gene functions for osteosarcoma (OS) by integrating network-based method with guilt by association (GBA) principle (called as network-based gene function inference approach) based on gene ontology (GO) data and gene expression profile. To begin with, differentially expressed genes (DEGs) were extracted using linear models for microarray data (LIMMA) package. Then, construction of differential co-expression network (DCN) relying on DEGs was implemented, and sub-DCN was identified using Spearman correlation coefficient (SCC). Subsequently, GO annotations for OS were collected according to known confirmed database and DEGs. Ultimately, gene functions were predicted by means of GBA principle based on the area under the curve (AUC) for GO terms, and we determined GO terms with AUC > 0.7 as the optimal gene functions for OS. Totally, 123 DEGs and 137 GO terms were obtained for further analysis. A DCN was constructed, which included 123 DEGs and 7503 interactions. A total of 105 GO terms were identified when the threshold was set as AUC > 0.5, which had a good classification performance. Among these 105 GO terms, 2 functions had the AUC > 0.7 and were determined as the optimal gene functions including angiogenesis (AUC =0.767) and regulation of immune system process (AUC =0.710). These gene functions appear to have potential for early detection and clinical treatment of OS in the future.

1. Introduction

Osteosarcoma (OS) is the most commonly diagnosed histological form of primary bone tumor with high morbidity, and it is mainly prevalent in teenagers and young people [1]. Pain is the most common early symptom of OS and can cause fracture of the affected bone. Currently, multiple therapeutic strategies for OS, for example, surgical resection, chemotherapy and radiotherapy, have significantly improved the prognosis of patients with OS [2]. However, the overall survival rates rarely exceed 60–65% [3]. Moreover, a significant portion of OS patients develop metastasis even after curative resection of the primary tumor. There is still a long way to go for the management of OS [4]. Therefore, with an attempt to continue to make progress in the diagnosis and management of OS, identification of sensitive and specific minimally invasive signatures is one of the most important challenges.

Genetic aberration has been demonstrated as an important factor that may play significant roles in OS pathogenesis. For instance, Zhu and colleagues [5] have reported that *SOX9* is over-expressed in OS tissues compared with controls. Moreover, it is indicated that *FOXM1* is up-regulated, which suggests *FOXM1* might be an valuable bio-signature for OS [6]. Nevertheless, these genes do not treat the OS efficiently or selectively, because molecules frequently don't work individually, yet co-operate with the other genes. Additionally, genetic factors can disturb the protein levels, thereby in turn perturb the molecule interactions. Network is characterized by the complicated interactions and the complex interwoven relationships that control cellular functions [7]. Hence, understanding the networks will be beneficial to provide new insights to explore the molecular pathogenesis of OS. The concept of differential co-expression network (DCN) has been employed to the studies of OS, due to statistical confidence of single connections, overlap with protein interaction, and mathematical convenience [8]. In addition, improving our knowledge of gene function in uncharacterized genes is a major task [9]. Remarkably, gene interactions can be applied to deduce functional relationships based on a principle known as "guilt by association" (GBA) [10]. GBA has been indicated to predict gene function in various types of biological networks, for example, gene co-expression network [11].

Thus, in our work, we integrated network-based method with GBA principle (called as network-based gene function inference approach) to further identify the optimal gene functions for OS using gene ontology (GO) data and gene expression profile. To begin with, differentially expressed genes (DEGs) were extracted using linear models for microarray data (LIMMA) package. Then, construction of DCN based on DEGs

E-mail address: chenxr689@163.com.

http://dx.doi.org/10.1016/j.jbo.2017.04.003
Received 15 March 2017; Received in revised form 5 April 2017; Accepted 6 April 2017
Available online 08 April 2017

was implemented, and sub-DCN was identified using Spearman correlation coefficient (SCC). Subsequently, GO annotations for OS were collected according to known confirmed database and DEGs. Ultimately, gene functions were predicted by means of GBA principle based on the area under the curve (AUC) for GO terms, and we determined GO terms with AUC > 0.7 as the optimal gene functions for OS. These gene functions appear to have potential for early detection and clinical treatment of OS in the future.

2. Materials and methods

2.1. Gene expression profile and data pre-treatment

OS-related microarray data (accession number E-GEOD-36001) [12] were downloaded from ArrayExpress database based on the platform of A-MEXP-930-Illumina Human-6 v2 Expression BeadChip. There were 19 OS samples and 6 normal samples in E-GEOD-36001. Prior to analysis, we firstly pre-processed the microarray profile of E-GEOD-36001. Specifically, robust multi-array average (RMA) was used to perform background correction [13]. Quartile algorithm was utilized to implement quartile normalization [14], following by perfect match (PM)/mismatch (MM) correction using microarray suite (MAS) 5.0 package [15]. Ultimately, the data on probe levels were converted into gene symbols relying on annotate package [16]. Overall, 19,027 genes were reserved for further exploitation.

2.2. Identification of DEGs

The LIMMA package [17] in R language and an empirical Bayes framework were applied in our analysis to achieve DEGs between OS and normal samples. A *t*-test was conducted, and the multiple test was applied to correct the raw P values using the Benjamini & Hochberg [18] method based on false discovery rate (FDR). DEGs were extracted on the basis of FDR < 0.05 and |log fold change (FC)| ≥2.

2.3. Generation of DCN

Cytoscape (http://cytoscape.org/) is an open source software which integrates bio-molecular interactions with high-throughput expression data as well as other molecular states into a unified conceptual network [19]. Therefore, we inputted DEGs into the Cytoscape software to visualize the DCN. Furthermore, with the goal of assessing the co-expressed strength of each interaction in the DCN, SCC was used in our study. As we all know, SCC is used to measure the co-expression probability of two variables by assessing the strength of association of two co-expressed variables and it ranges from −1 to 1 inclusive [20,21]. The SCC absolute value of one interaction was determined as the weight value of the corresponding edge, and the higher the weight value was, the more relevant the interaction was related to the disease. Thus, we selected the edges with weight values higher than 0.8 to construct the sub-DCN. In order to display the sub-DCN more vividly, Cytoscape software was applied.

2.4. GO annotation for DEGs

GO annotation has been broadly utilized as functional enrichment studies for large-scale genes [22]. In our study, human GO annotations comprised of 19,003 GO terms covering 18,402 genes were retrieved from the GO consortium (http://geneontology.org/). In an attempt to obtain stable performance, we filtered for the GO annotations on gene size such that each remaining GO term had associated genes > 20, and a total of 1313 GO terms were left to be used in our analysis. Next, DEGs identified above were mapped to the 1313 GO annotations. Finally, the GO slim set was obtained, consisting of 123 DEGs and 137 GO terms.

2.5. Identifying gene functions based on "GBA" prediction

Gene networks have been widely used in gene function prediction algorithms, many based on complex extensions of the "GBA" principle. As demonstrated here, GBA prediction approach was utilized to predict significant gene functions in OS progression. For GBA method, we used three-fold cross-validation to extract a ranked list scoring genes in DCN as to how they belonged in the known GO terms. The sum of co-expression values between the training set (co-expression) and the candidate gene was divided by the sum of co-expression values between the genes outside the training set and the candidate gene to analyze degree of candidacy. In detail, for each gene *i* in the DCN, all other neighbored genes of gene i were mapped to each GO category K, and the multifunctionality (MF) score for each gene i within the K-GO term based on the following equation:

$$\mathbf{MF(gene_i) = \sum_{i|gene_A \;\; \in GO_K} 1/N_{in_K} {*} N_{out_K}}$$

In this formula, N_{in_i}stood for the number of genes in GO term i, and N_{out_i} denoted the count of genes outside GO category i.

Then, based on support vector machine (SVM), AUC for each GO group K was calculated, and the mean AUC across all GO terms was determined. Thus, the AUC scores were ordered from the highest to the lowest, the ranks of GO terms were ordered oppositely. The AUC of 0.5 stands for classification at chance levels, while the AUC of 1.0 denotes a perfect classification. In the literature about the gene function prediction, the AUC greater than 0.7 are regarded good [23]. In our study, GO terms with AUC > 0.7 were identified and regarded as the optimal gene functions.

3. Results

3.1. DCN construction

Before DCN generation, we firstly identified DEGs between OS and normal samples using *t*-test. Based on FDR < 0.05 and |log FC = ≥ 2, a total of 123 DEGs were identified. The top 20 DEGs were shown in Table 1. The most significant DEGs were *HOXB7*, *RHPN2*, *SRGN*, *FOXF2*, and *PLVAP*.

In order to further explore the biological activities of DEGs, we constructed a DCN using the above-identified 123 DEGs. In this DCN, there were 123 nodes and 7503 interactions. Under the DCN, node degree was more than just an statistic about a gene, and significantly,

Table 1
List of the top 20 differentially expressed genes (DEGs).

Genes	|log (fold change)|	False discovery rate (FDR)
HOXB7	2.085898	1.33E−07
RHPN2	2.147406	6.32E−07
SRGN	5.174341	3.34E−06
FOXF2	2.318869	4.51E−06
PLVAP	3.404251	1.98E−05
COX7A1	2.988735	5.41E−05
APOE	3.399852	6.21E−05
SPINT2	2.301193	6.65E−05
LXN	2.523381	8.93E−05
TNFRSF1B	2.427194	1.10E−04
VAMP8	3.494650	1.61E−04
CBS	2.868625	1.76E−04
HCLS1	2.686705	1.81E−04
PHGDH	2.362045	2.15E−04
GIMAP7	2.258177	2.74E−04
C1QC	2.885076	2.76E−04
HBB	5.191561	2.86E−04
C1QA	3.521570	2.93E−04
TYROBP	3.511873	3.17E−04
C1QB	3.184835	3.17E−04