Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Identification of transcription factors that may reprogram lung adenocarcinoma

Chenglin Liu [a,1], Yu-Hang Zhang [b,1], Tao Huang [b,*], Yudong Cai [c,*]

[a] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China
[b] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China
[c] School of Life Sciences, Shanghai University, Shanghai 200444, People's Republic of China

ARTICLE INFO

Article history:
Received 19 December 2016
Received in revised form 19 January 2017
Accepted 22 March 2017

Keywords:
Lung adenocarcinoma
Transcription factor
Tumorigenesis
Greedy algorithm

ABSTRACT

Background: Lung adenocarcinoma is one of most threatening disease to human health. Although many efforts have been devoted to its genetic study, few researches have been focused on the transcription factors which regulate tumor initiation and progression by affecting multiple downstream gene transcription. It is proved that proper transcription factors may mediate the direct reprogramming of cancer cells, and reverse the tumorigenesis on the epigenetic and transcription levels.
Methods: In this paper, a computational method is proposed to identify the core transcription factors that can regulate as many as possible lung adenocarcinoma associated genes with as little as possible redundancy. A greedy strategy is applied to find the smallest collection of transcription factors that can cover the differentially expressed genes by its downstream targets. The optimal subset which is mostly enriched in the differentially expressed genes is then selected.
Results: Seven core transcription factors (MCM4, VWF, ECT2, RBMS3, LIMCH1, MYBL2 and FBXL7) are detected, and have been reported to contribute to tumorigenesis of lung adenocarcinoma. The identification of the transcription factors provides a new insight into its oncogenic role in tumor initiation and progression, and benefits the discovery of functional core set that may reverse malignant transformation and reprogram cancer cells.

1. Introduction

Cancer as a systematic disease, has been widely regarded as one of the newly rising threatens to human health [1–3]. It can start anywhere in the body when malignant cells grow out of control and start to eliminate normal cells, causing the classified metabolic abnormalities in the body [4]. Based on the statistics of WHO in 2012, cancers turn out to be the leading causes of morbidity and mortality around the world, inducing about 14 million new cases and 8.2 million cancer related deaths. The number of new cases may further rise by about 70% within the following two decades, reaching 22 million new cases a year [5]. As we all know, cancer is a generic term for a large group of systematic diseases, involving different parts of the body [6,7]. Among different subtypes of cancer, lung adenocarcinoma, deriving from the bronchial epithelium cell has been confirmed to be one of the top killers for human health [8]. Based on the statistics from WHO in 2012, 1.59 million deaths around the world can be attributed to lung cancer [9–11]. Furthermore, among lung cancer, lung adenocarcinoma accounts for more than 40% cases, implying that lung adenocarcinoma may be one of the most threatening diseases to human health [5].

Decades ago, cancer has already been confirmed to be induced by the stimulation of both genetic and environmental factors. Specifically for lung adenocarcinoma, genetic background has been confirmed to play an irreplaceable role during the tumorigenesis [11]. Based on sequencing identification and experimental validation, the specific genetic background of lung adenocarcinoma has been preliminarily revealed. Genes like *EGFR*, *KRAS*, *MET*, *LKB1*, *BRAF*, *PIK3CA*, *ALK*, *RET* and *ROS1* have all been confirmed to contribute to the initiation, progression or metastasis processes of lung adenocarcinoma, implying the complicated characteristic genetic background of such disease [12]. However, during the identification of lung adenocarcinoma associated genes and proteins, a group of functional regulatory proteins, the transcription factors have been consciously or unconsciously overlooked and not preferred. As we all know, transcription factor turns out to be a protein that binds to

* Corresponding authors.
E-mail addresses: tocliu@126.com (C. Liu), zhangyh825@163.com (Y.-H. Zhang), tohuangtao@126.com (T. Huang), cai_yud@126.com (Y. Cai).
[1] Co-first authors.

http://dx.doi.org/10.1016/j.artmed.2017.03.010
0933-3657/

specific DNA sequences and contributes to the regulation of various downstream gene transcription [13]. Therefore, transcription factors may directly regulate the functions of crucial genes in human bodies physically or psychologically.

Transcription factors have been widely reported to contribute to the regulation of biological processes. With the deepening of the understanding on biological systems, people began to not only study on the function of transcription factors, but also tried to introduce transcription factors as biological tools to control or even reverse the biological processes. In 2006, Japanese scientist, Shinya Yamanaka presented a timepiece revolutionary technology producing induced pluripotent stem cells (iPS cells), reversing the development processes of differentiated cells from adults. By injecting four specific transcription factor: Oct4, Sox2, c-Myc, Klf4, Shinya Yamanaka successfully transformed the differentiated adult human dermal fibroblasts into human iPS cells, which are quite similar with human embryonic stem (ES) cells [14], Such historically significant achievement validated that it's possible to reverse the cellular transformation, like differentiation/dedifferentiation during development by transcription factor interferes. As we all know, similar with the reversion of development processes, tumor genesis also involve various dedifferentiation processes regulated by functional transcription factors [15,16]. Considering the similarity of development process and tumor genesis, transcription factors may definitely play an irreplaceable role during the dedifferentiation and proliferation of tumor cells.

As we have analyzed above, proper transcription factors may also regulate the direct reprogramming of malignant cells which further contribute to the initiation and progression of tumor tissues during tumor genesis, just like in iPS system [17–19]. According to recent publications, the key contribution of transcription factors and the core regulatory transcription factors that contribute to tumor initiation and progression have not been fully revealed. Only a few single transcription factors have been identified to involve in tumor genesis, such as PAX-7, FOXA2 and TTF-1 in lung cancer [20–22].Specific transcription factors like POU3F2, SOX2, SALL2 and OLIG2 have also been reported to directly contribute to the malignant proliferation of glioma cells, validating the core role of functional transcription factors during tumor genesis in multiple tumor subtypes [23]. Founded on the functional study of transcription factors in cancers, recent publications further confirmed that proper application of transcription factors may even reverse the cancer progression, validating the crucial upstream regulatory role of transcription factors [24,25]. In lung cancer, it has been reported that direct reprogramming of cancer cells by optimal transcription factors (OCT3/4, KLF4, c-MYC and SOX2) may reverse the tumor genesis at least on the epigenetic and transcription levels, validating the reversible regulatory characteristics of transcription factors during tumor genesis [25]. Significant as transcription factors are in cancer, however, few detailed transcription factors like OCT3/4 as we mentioned above have been identified. Since lung cancers turn out to be one of the most threatening diseases to human health and transcription regulations have been confirmed to be quite significant, it's quite necessary and urgent to compensate the lack of study in this field, providing optimal transcription regulatory components for further reversion of tumorigenesis in multiple tumor subtypes including lung cancer.

Here, we take lung adenocarcinoma as a model, and construct a computational framework to predict the optimal transcription factors for the direct reprogramming. The method is based on the assumption that the core regulatory transcription factors can regulate as many as possible lung adenocarcinoma associated genes (such as EGFR, MLL2) with as little as possible redundancy [26,27]. Firstly, a greedy strategy is employed to find the smallest collection of transcription factors that the union of their downstream genes can cover the differentially expressed genes. Then, the core transcription factors are selected that their downstream genes are most enriched in the differentially expressed genes. Seven core transcription factors (MCM4, VWF, ECT2, RBMS3, LIMCH1, MYBL2 and FBXL7) are detected after that, all of which have been confirmed to regulate lung adenocarcinoma associated genes according to published literature. Such transcription factors may definitely further regulate the initiation and progression of lung adenocarcinoma, and the abnormal function of such transcription factors may promote the tumorigenesis. The optimal regulation of such transcription factors, on the contrary, may reverse the cancerous process and suppress the progression of tumor. All in all, for the first time, we screened and identified the core regulatory tumor associated transcription factors according to their specific functions, their downstream regulated genes, not just their respective mutation or expression variants. The presentation and application of our newly presented computational methods may develop a functional tool to identify tumor, including lung adenocarcinoma associated transcription factors, providing a new insight into the oncogenic role of transcription factors during tumor initiation and progression, and benefiting the identification of core functional transcription factors that may reverse malignant transformation.

2. Material and method

2.1. Datasets

The differentially expressed genes of lung adenocarcinoma comparing to their respective normal tissues were obtained from ONCOMINE database (www.oncomine.org) [28]. ONCOMINE database is a widely used platform for bioinformatics discovery from genome-wide microarray expression analysis. The platform has included more than 700 independent datasets and performed differentially expression analyses comparing various types of cancer with respective normal tissues. Ten published studies were integrated in the database covering the profiling of 1133 samples, as shown in Table 1. The differentially expressed genes were deter-

Table 1
The statistics of the selected datasets in ONCOMINE database for the current study.

Dataset Name	Resources (Citation or GEO accession number)	#Normal	#Cancer	#DEG
Beer et al.	Nat Med 2002 [65]	10	86	532
Bhattacharjee et al.	Proc Natl Acad Sci 2001 [66]	17	139	858
Garber et al.	GSE3398 [67]	5	40	1094
Hou et al.	GSE19188 [68]	65	45	1956
Landi et al.	GSE10072 [69]	49	58	1262
Okayama et al.	GSE31210 [70]	20	226	1956
Selamat et al.	GSE32863 [71]	58	58	1926
Stearman et al.	GSE2514 [72]	19	20	860
Su et al.	GSE7670 [73]	30	27	1262
Yamagata et al.	Clin Cancer Res 2003 [74]	3	9	244

#: Number. DEG: differentially expressed genes, including the top 5% over-expressed and top 5% under-expressed genes with FDR adjusted p-values less than 0.05.

Please cite this article in press as: Liu C, et al. Identification of transcription factors that may reprogram lung adenocarcinoma. Artif Intell Med (2017), http://dx.doi.org/10.1016/j.artmed.2017.03.010