Surgical Oncology 23 (2014) 126–131

Contents lists available at ScienceDirect

Surgical Oncology

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

Review

A microRNA gene signature for identification of lung cancer

Zhiqiang Xue [a, 1], Jiaxin Wen [a, 1], Xiangyang Chu [a], Xinying Xue [b, c, *, 1]

[a] Department of Thoracic Surgery, Chinese PLA General Hospital, Beijing, China
[b] Department of Special Medical Treatment – Respiratory Disease, Beijing Shijitan Hospital, Capital Medical University, Beijing, China
[c] Department of Respiratory Diseases of Chinese PLA General Hospital, Beijing, China

ARTICLE INFO

Article history:
Accepted 27 April 2014

Keywords:
MicroRNA
Gene
Identification cancer
Lung cancer
Gene signature

ABSTRACT

Background: Lung cancer is the leading cause of cancer deaths worldwide, compounded by late diagnosis. MicroRNAs (miRNA) are recently discovered short, noncoding genes that play essential roles in tissue differentiation during normal development and tumorigenesis. miRNA profiles across all histologic grades can provide a reliable and standardized method for the identification of lung cancer.
Methods: A microRNA lung cancer dataset was analyzed. Differentially expressed microRNAs were obtained post-normalization of data using *t*-test ($p < 0.01$). The data for differentially expressed microRNAs were processed using K-nearest neighbors classification method to obtain unique miRNAs expression patterns. The predicted mRNA targets were identified using TargetScan and the molecular functions associated with the predicted targets were retrieved from the Gene Ontology Consortium and represented using GO IDs in a directed acyclic graph.
Results: The results indicate that lung cancer samples can be classified using a small panel of 19 unique microRNAs (8 down-regulated and 11 up-regulated) with over 85% classification accuracy. Furthermore, using classical enrichment analysis, this study identified 66 molecular function groups which are potentially the functional signaling pathways altered by these differentially expressed microRNAs.
Conclusions: We identified a microRNA gene signature representative of functioning as a diagnostic biomarker for lung cancer. These findings can potentially form the basis for the development of a standardized diagnostic assay that can be used for early diagnosis of lung cancer equally well from resection specimens and cytology samples.

Contents

* Corresponding author. Department of Special Medical Treatment – Respiratory Disease, Beijing Shijitan Hospital, Capital Medical University, 10 Tieyi Rd., Beijing, China. Tel./fax: +86 010 63925588.
E-mail address: xinyingxue2010@163.com (X. Xue).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.suronc.2014.04.003
0960-7404/

Introduction

Despite rapid advances in drug development and surgical procedures, lung cancer is the leading cause of cancer death worldwide in both men and women [1], with an estimated 1.4 million deaths each year [2]. An estimated 1.6 million new cases of lung cancer were diagnosed worldwide in 2008, accounting for about 13-percent of total cancer diagnoses [1]. It is estimated that approximately 15% of lung cancer patients will have a 5-year survival from the time of diagnosis [3]. The alarming nature of these numbers is further highlighted by the fact that lung cancer death is greater than the next three most common cancers combined, inclusive of colon, breast, and prostate cancer. This is largely due to the fact that lung cancer is not usually diagnosed until advanced stage [4].

The gold standard of lung cancer classification was routine histopathology; however, it had its limitations [5–7]. This problem gets further magnified in cases where small amount of diagnostic samples are available making it difficult to obtain adequate number of tumor cells with desired tissue architecture [6–9]. Though immunohistochemistry achieved better accuracy of classification it had the limitations varying from technical inconsistencies, tumor heterogeneity, to lack of specificity and sensititvity of each individual marker [8–12]. This highlights the need for a more reliable and non-invasive screening and diagnostic tool for lung cancer.

MicroRNAs (miRNAs) are 21–23 nucleotides noncoding genes that have been extensively cataloged in post transcriptional regulation of gene expression [13] during normal development and tumorigenesis [14–17]. Since miRNA expression is temporally regulated, the expression profiles mimic the developmental lineage and differentiation pattern of the tumors, allowing miRNA expression level to identify poorly differentiated tumors, which is impossible just by looking at mRNA profiles [18].

Hence, miRNA profiling is imperative for understanding the overall impact of post transcriptional regulation of gene expression in lung cancer cells, which in turn will help identify cognate therapeutic targets and curative measures [19]. What makes miRNA profiling inherently complex is that the same gene may be targeted by more than one microRNA. However, detailed analysis of miRNA and cognate targets can be used to develop microRNA signatures that can be used for lung cancer classification. Hierarchical clustering of gene expression data obtained from miRNA profiling was successfully used to classify cancers based on their tissue origin and developmental pattern [8]. On the contrary, clustering done for gene expression data obtained from mRNA profiling did not have similar classification power, highlighting the significant impact of successful miRNA profiling and analyses. The purpose of the current study were to: (*a*) hierarchical gene clustering analysis of previously obtained dataset to identify putative miRNA signature in lung cancer; and (*b*) establish the overall sensitivity and specificity of miRNA profiling in lung cancer classification.

Materials and methods

Dataset

A microRNA dataset consisting of normal and lung cancer data was imported from European Bioinformatics Institute's (EBI's) database dated 2005-10-31 and analyzed. The dataset was collected from microRNA expression profiles using microarray analysis. This dataset comprises of 246 files, equally divided among normal lung and human lung cancer data. Each of 246 files has 25 columns of information and 374 features. For the current analysis only the Reporter name and Signal Noise Ratio columns were extracted from each file for further analysis. The Reporter name column included the identifiers of all microRNAs which were probed in the microarray experiments and the Signal Noise Ratio column contained the expression levels of the microRNAs. To identify a panel of differentially expressed microRNAs, a new dataset containing all the related features was generated.

microRNA array data normalization

To ensure the results from the experiments can be compared, the expression profiles (columns of data) were normalized using the following formula

$X'_{ij} = [(X_{ij} - \hat{X}_j)/\sigma_j], \quad i = 1, \ldots, 374 : j = 1, \ldots, 246$ where, X_{ij} is the expression level of microRNA i in sample j (column j); $\hat{X}_j$ is the geometric mean of the sample j; σ_j is the standard deviation of the sample j; and, X'_{ij} is the normalized expression level of microRNA i in sample j.

Identifying unique miRNA expression pattern in cancer samples

Overall flow of the current analysis is summarized in Fig. 1. The input data for all 374 microRNAs from each of the 246 samples were normalized. *t*-test was used to process the normalized data to identify differentially expressed microRNAs ($p < 0.01$). The command used to perform the *t* test is detailed in the Supplementary Data File.

Classification of sample sets using differentially expressed miRNAs

Using 'R', 50 test cases were formed by randomly choosing 12 columns (about 10% of the samples) from each group to make a test set; the remaining columns constituted the training set in each test case. The data for differentially expressed microRNAs thus obtained were further processed using K-nearest neighbors (KNN) classification method [20] to obtain unique miRNAs expression patterns in the lung cancer samples. Since, the first twelve columns actually were noncancerous data and remaining twelve columns were cancerous, the result was compared with the expected result and the accuracy percentage was calculated. For the same pair of testing-training set, predicted outputs were calculated for $K = 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37$ and 39. The same procedure was carried out for all 57 (the total number of differentially expressed miRNAs in lung cancer identified through the *t* test) pairs of randomly selected testing set-training set. This procedure was conducted for data of top 5, 10, 20, 40 and 57 microRNAs. Data of the top 5, top 10, top 20, top 40 and top 57 microRNAs were analyzed separately to study the degree of distinctness of data in each one of them. Average values of the 50 test cases were calculated. The standard deviations of the outputs and standard errors were also calculated. The unique microRNAs were subsequently separated into up-regulated and down-regulated categories compared to normal lung expression profile.

From the overall result, it was concluded that the distinctness of a data belonging to one class from those belonging to another class was very strong in our dataset. Since, the result obtained while