Computers in Biology and Medicine 42 (2012) 624–630

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

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

Screening for cancer associated MiRNAs through co-gene, co-function and co-pathway analysis

Xue Xiao [a], Dongguo Li [a], Lei Gao [a], Xia Li [a,b,*], Qianghu Wang [b], Shaojun Zhang [b], Zhicheng Liu [a]

[a] School of Biomedical Engineering, Capital Medical University, Beijing 100069, China
[b] College of Bioinformatics Science and Technology, and Bio-pharmaceutical Key Laboratory of Heilongjiang Province and State, Harbin Medical University, Harbin, Heilongjiang 150086, China

ARTICLE INFO

Article history:
Received 4 July 2011
Accepted 25 February 2012

Keywords:
MicroRNA
Cancer
Enrichment analysis
Semantic similarity

ABSTRACT

MicroRNAs (miRNAs) though present themselves as a group of non-coding small RNAs play critical roles in many biological and pathological processes. Among which the regulation of human cancer is one of the most excited potentiality. The goal of this study is to obtain miRNAs robustly associated with cancer by screening all of the possible miRNAs/cancer pairs in three consecutive steps. First, in co-gene analysis, gene set enrichment analysis is carried out for all miRNA/cancer pairs. Second, in co-function analysis, information theoretic similarity on GO is calculated for miRNA/cancer pairs screened from the former step. Third, in co-pathway analysis, pathway enrichment analysis is performed for miRNA/cancer pairs screened from the second step. In this study, we totally included 776 miRNAs and 25 cancer types. As a result, 94 miRNAs were identified with robust association with 17 types of cancer. Meanwhile, 83 pathways with relevance to both miRNAs and cancer were also singled out. This framework provides an effective way to narrow down miRNAs for cancer and to pinpoint corresponding pathways.

1. Introduction

As a group of endogenous non-coding small RNA molecules, microRNAs (miRNAs) are usually involved in negative post-transcriptional regulation of gene expression [1]. The genes that are regulated by miRNAs are known as 'target genes'. Through the regulation of target genes, miRNAs can affect many fundamental physiological processes such as cell proliferation, differentiation, and apoptosis [2,3]. MicroRNAs are further implicated in many kinds of disease, such as cancer [3–5]. A piece of early solid evidence supporting a tight association between miRNAs and cancer was presented in 2002 [4]. Up till now, cumulative evidence shows that miRNAs can impact cancer through regulation of cancer genes [6–9]. Moreover, miRNAs have become potential biomarkers for cancer diagnosis and prognosis [10–12].

Today, we have hundreds of miRNAs and dozens of cancer types. Meanwhile, there are well-designed databases offering miRNA target genes and cancer genes. These resources enable us to make a comprehensive investigation of associations between miRNAs and different cancer types with computational methods. Gene set enrichment analysis was applied to find out cancer associated miRNAs [13]. However, due to the high false positive rate of miRNA target gene prediction, the association built on gene set enrichment alone has weak reliability. In addition, rather than working independently, a gene collaborates with other genes to function. Through collaboration, different genes may be involved in similar functions. Therefore, it makes more sense to take biological background of genes into consideration. Here, we consider that if a miRNA and a type of cancer not only have a significant overlap in gene set but also have similar biological functions, then the association between them will be more reliable. Furthermore, if the miRNA and the type of cancer are implicated in the same pathways, then not only the association between them will be more robust but also the molecular mechanisms of the association can manifest themselves in these pathways. Based on the thoughts, we propose a framework to achieve cancer associated miRNAs (CAMs) and relevant pathways in three consecutive steps: co-gene analysis, co-function analysis and co-pathway analysis.

2. Materials and methods

2.1. Databases of cancer genes and miRNA target genes

Cancer genes were acquired from the National Cancer Institute, which included around 1622 cancer genes, (http://ncicb.nci.nih.gov/NCICB/projects/cgdcp, Phrase 6). The raw

* Correspondence to: School of Biomedical Engineering, Capital Medical University, No. 10 Xitoutiao, Youanmenwai, Beijing 100069, China.
Tel.: +86 10 83911559, Fax: +86 10 83911560.
E-mail address: lixia6@yahoo.com (X. Li).

0010-4825/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2012.02.008

data contained the information of gene symbols, matched disease terms and evidence code. With keywords of different cancer types, genes were extracted and sorted into 35 cancer gene sets. To improve the reliability, only genes verified in experiment (evidence code: EV-EXP) were singled out to join the further study. Cancer gene sets containing less than 10 genes were discarded. In the end, 25 cancer gene sets and 802 cancer genes were used in our study.

The human target genes were downloaded from seven databases: RNAhybrid [14], DIANA-microT [15], RNA22 [16], miRBase [17], miRanda [18], PicTar [19], and TargetScan [20]. Considering that every database suffers from a high false positive rate of target prediction, we collected MiRNA/target matches cross-validated by at least two databases to increase the reliability of the data. In fact, the similar problem in protein–protein interaction (PPI) data has been solved also by integrating at least two PPI databases[21]. In total, the study included 776 human miRNAs, 14622 target genes and 283905 miRNA/target matches. Because of the intersection between the databases, a miRNA had on average 366 target genes.

2.2. The screening framework

The screening framework consumes miRNA target gene sets and cancer gene sets as input to produce CAMs and relevant pathways as output. The whole process is presented schematically in Fig. 1.

2.2.1. Co-gene analysis

Hypergeometric distribution was utilized as the method to calculate the significance of the overlap between a cancer gene set (CG) and a miRNA target gene set (MG).

$$P = 1 - \sum_{i=0}^{H-1} \frac{\binom{M}{i}\ \binom{N-M}{K-i}}{\binom{N}{K}}$$

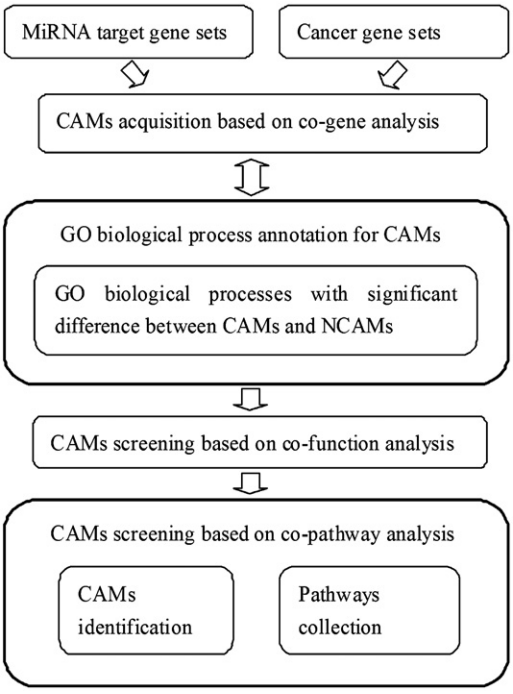

Fig. 1. A schematic overview of the CAM screening process. The framework generates reliable miRNA/cancer pairs, each of which consists of a CAM and a cancer type, as well as the pathways related with both CAM and cancer.

where N is the size of background gene set, which consisted of 27842 human genome genes (obtained from NCBI Entrez Gene, version 2007). M is the size of MG. K is the size of CG. H is the size of the overlap between MG and CG. MiRNA/cancer pair would be reserved when the P-value was less than 0.001.

2.2.2. GO biological process annotation

To ensure that a screening step was in the right direction, we measured the Gene Ontology (GO) based functional difference between cancer associated miRNAs (CAMs) and non-cancer associated miRNAs (NCAMs) as follows: 1) EASE 2.0[22] was employed to perform GO annotation for CAMs and NCAMs. For each miRNA, its target gene list and the human genome gene list were submitted. The output file contained GO terms and corresponding EASEscores (an adjusted Fisher exact probability). 2) The GO biological processes with EASEscore less than 0.05 were selected. 3) For each of the selected GO biological processes, the number of annotated miRNAs was respectively calculated for CAMs and NCAMs. Subsequently, chi square test was applied with R (http://www.r-project.org/). 4) GO biological processes with significant difference (p-value < 0.001) in annotation rate between CAMs and NCAMs were collected.

2.2.3. Co-function analysis

In co-function analysis, MG and CG are first annotated in GO. Assuming that the numbers of annotated GO terms are X and Y respectively, similarity between two GO terms can be calculated as follows [23,24]:

$$s(t1,t2) = \frac{2IC_{ms}(t1,t2)}{IC(t1)+IC(t2)}$$

where $t1$ represents a annotated GO term of MG. $t2$ represents a annotated GO term of CG. IC is the information content of a GO term. IC_{ms} denotes the information content of the most informative common ancestor of two GO terms.

Calculation of s-values between Y GO terms of CG and X GO terms of MG generates a $Y \times X$ similarity matrix S. Information theoretic GO similarity between CG and MG can be computed as follows:

$$\text{rowScore} = 1/Y \sum_{i=1}^{Y} \max_{1 \le j \le X} S_{ij}$$

$$\text{columnScore} = 1/X \sum_{j=1}^{X} \max_{1 \le i \le Y} S_{ij}$$

$$\text{Sim}(c,m) = 1/2(\text{rowScore} + \text{columnScore})$$

In our study, if Sim-value was higher than 0.7, which was a significantly high value in random case (p-value < 0.0001), then CG and MG were deemed to be highly similar in biological function. Accordingly, the miRNA/cancer pair would be reserved.

2.2.4. Co-pathway analysis

In co-pathway analysis, KEGG [25] pathway enrichment analysis was performed respectively for MG and CG using hypergeometric distribution. If MG and CG enriched (p-value < 0.001) in at least one identical pathway, then the corresponding miRNA/cancer pair would be reserved. At the mean time, the pathways enriched with both MG and CG would be collected as shared pathways.

3. Results

3.1. Screening with co-gene analysis

To elucidate the screening procedure, 776 human miRNAs and 25 types of human cancer were included, which were randomly