Computational Biology and Chemistry 44 (2013) 15–21

Contents lists available at SciVerse ScienceDirect

Computational Biology and Chemistry

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

Analysis of the NCI-60 dataset for cancer-related microRNA and mRNA using expression profiles

Chia-Wei Weng[a], Shan-Chih Lee[b], Yu-Liang Lee[a], Ka-Lok Ng[a,*]

[a] Department of Biomedical Informatics, Asia University, 500 Lioufeng Road, Wufeng Shiang, Taichung 41354, Taiwan
[b] School of Medical Imaging and Radiological Sciences, Chung Shan Medical University, Taichung 40201, Taiwan

ARTICLE INFO

Article history:
Received 10 October 2012
Accepted 12 February 2013

Keywords:
MicroRNA
Expression profile
Oncogene
Tumor suppressor gene
Pearson correlation
Spearman rank correlation
CpG island
MicroRNA cluster
Fragile sites

ABSTRACT

Background: Recent studies have indicated that microRNA (miRNA) may play an oncogenic or tumor suppressor role in human cancer. To study the regulatory role of miRNAs in tumorigenesis, an integrated platform has been set up to provide a user friendly interface for query. The main advantage of the present platform is that all the miRNA target genes' information and disease records are drawn from experimentally verified or high confidence records.
Results: MiRNA target gene results are annotated with reference to the disease gene as well as the pathway database. The correlation strength between miRNA and target gene expression profile is quantified by computing the correlation coefficient using the NCI-60 expression profiling data. Comprehensive analysis of the NCI-60 data found that the cumulative percentage of negative correlation coefficients for cleavage regulation is slightly higher than its positive counterpart; which indicated that the mRNA degradation mechanism is slightly dominant. In addition, the RNAHybrid and TargetScans scores are computed which potentially served as quantitative estimators for miRNA–mRNA binding events.

Three scores are defined for each miRNA–mRNA pair, which are based on the disease gene and pathway information. These three scores allow user to sort out high confidence cancer-related miRNA–mRNA pairs.

Statistical tests were applied to investigate the relations of three chromosomal features, i.e., CpG island, fragile site, and miRNA cluster, with cancer-related miRNAs. A web-based interface has been set up for query, which can be accessed at: http://ppi.bioinfo.asia.edu.tw/mirna_target/
Conclusions: The main advantage of the present platform on miRNA–mRNA targeting information is that all the target genes' information and disease records are experimentally verified. Although this may limit the number of miRNA–mRNA relationships, the results provided here are more solid and have fewer false positive events. Certain novel cancer-related miRNA–mRNA pairs are identified and confirmed in the literature. Fisher's exact test suggests that CpG island and fragile site associated miRNAs tend to associate with cancer formation. In summary, the present platform provides an easy means of investigating cancer-related miRNAs.

1. Introduction

MicroRNAs (miRNAs) are a class of small non-coding RNAs that bind to their target mRNA sequence in the 3′-untranslated region (3′-UTR), and induce either translation repression or mRNA degradation. Recent studies have indicated that miRNA could possibly play an important role in human cancer where miRNA targets oncogenes (OCG) or tumor suppressor genes (TSG) to regulate the gene expression (Chen, 2005; Garzon et al., 2006; Voorhoeve, 2010; Zhang et al., 2007). When miRNAs play an oncogenic role, they target TSG and lead to tumor formation. On the other hand, there are reports which show that miRNAs are repressed in human colon cancer (Michael et al., 2003), lung cancer (Takamizawa et al., 2004), breast cancer (Iorio et al., 2005), and renal cell cancer (Liu et al., 2010) compared with their respective tissues. In other words, if miRNA plays the role of tumor suppressor due to deregulation of miRNA, OCG may be over-expressed and induce tumor formation.

To study the role of miRNA in carcinogenesis, it is necessary to obtain high confidence miRNA target gene pairs. There are many tools available for the identification of miRNA target genes, such as miRanda (Enright et al., 2003), RNAHybrid (Kruger and Rehmsmeier, 2006), and TargetScans (http://genes.mit.edu/tscan/targetscanS2005.html) etc. A major problem regarding miRNA target genes prediction is that the prediction accuracy remains uncertain, there are reports indicating that the false positive rate could be as high as 50% for human (John et al., 2004), 24–39% and 22–31% when using miRanda

* Corresponding author. Tel.: +886 423394541; fax: +886 423320718.
E-mail addresses: jeff19811029@gmail.com (C.-W. Weng), sclee@csmu.edu.tw (S.-C. Lee), jackox.tw@gmail.com (Y.-L. Lee), ppiddi@gmail.com (K.-L. Ng).

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.02.001

(Bentwich, 2005), and TargetScan (Bentwich, 2005) respectively. If the miRNA–mRNA targeting part is uncertain, then any derived information is meaningless.

To address this problem, several publicly available cancer gene databases are utilized to provide disease genes information. The miRNA–mRNA target gene information is retrieved from the TarBase (Sethupathy et al., 2006) and miRTarBase data (Hsu et al., 2011). The main advantage of using TarBase and miRTarBase in constructing cancer-related miRNA is that all the target genes recorded by these two databases are experimentally verified. Therefore, cancer-related miRNA–mRNA pairs are identified by integrating the cancer gene, TarBase and miRTarBase data. From a biomedical point of view, these results are more reliable.

Given a miRNA–mRNA pair, the regulation strength is quantified by using both Pearson correlation coefficient (PCC) and Spearman rank correlation coefficient (SRC), in which the expression profiles are obtained from the NCI-60 (Blower et al., 2007; Hsu et al., 2011; Shankavaram et al., 2007) cancer cell line microarray data. The NCI-60 data sets include Affymetrix miRNA and mRNA expression profiles for nine cancer types. In this work, we provided miRNA target prediction scores from RNAHybrid and TargetScan algorithms as well. Although target gene prediction remains uncertain, these scores may serve as a reference for the PCC and SRC results.

To examine whether a miRNA–mRNA pair or a target gene recorded by TarBase or miRTarBase is involved in one or more cancer types or not, we conducted a keyword search. This search included miRNA ID, gene symbol, and disease disorder. Three sets of complementary data are adopted in such search; that is, disease-related miRNAs, disease genes and cancer pathways databases. Three scores are assigned to quantify the matching between Tarbase data and the three datasets. Essentially, there are three-fold benefits for integrating disease databases; (i) it provides a means of sorting out highly confident cancer-related miRNA–mRNA pairs; (ii) it suggests one or more tissue types for the study of effects of miRNA on carcinogenesis; and (iii) it can potentially identify novel miRNA–mRNA pairs involved in one or more cancer types among the nine cancer types from NCI-60.

Many reports have suggested that aberrant miRNA expression is associated with tumor progression and metastasis. Therefore, it is vital to understand the regulatory mechanism of miRNA expression. In this study, three types of chromosomal features are considered, that is, the presence of (i) CpG island; (ii) miRNA cluster; and (iii) fragile site, which are possibly linked to miRNA aberrant expression, and hence induce cancer formation.

It is reported that the presence of DNA methylation-associated silencing of miRNA is involved in human cancers. For example, the expression of the human miR-127 is specifically down regulated in multiple solid cancers. This miRNA is located within a CpG island, which is specifically hypermethylated in cancers (Deng et al., 2008; Lujambio and Esteller, 2007; Lujambio et al., 2007, 2008; Saito et al., 2006). Another work reported that four miRNAs located around CpG islands (i.e., miR-34b, miR-137, miR-193a, and miR-203) are silenced by DNA hypermethylation in colon cancer (Bandres et al., 2009; Toyota et al., 2008), hepatocellular (Furuta et al., 2010), and oral cancer (Kozaki et al., 2008).

A fragile site is a heritable region on a chromosome that tends to break or appear as a gap or constriction when the cell is exposed to certain chemical agents. Recent study has indicated that certain diseases are due to miRNAs located at human chromosome fragile sites (Calin et al., 2004). A similar finding was discovered in mouse (Sevignani et al., 2007). It is reported that 52.6% of miRNA genes are in human cancer-associated genomic regions or in fragile sites (Calin et al., 2004). Several investigations have reported that two miRNAs (i.e., miR-34a-1, miR-34a-2) located at fragile sites are shown to be involved in breast (Ma et al., 2007), and lung cancer (Bandi and Vassella, 2011).

It is known that miRNA genes are often form clusters. A miRNA cluster includes two or more miRNA genes. For instance, recent works report that the miR-17-92 cluster, composed of seven miRNAs and residing in intron 3 of the C13orf25 gene at 13q31.3, is markedly and frequently overexpressed in cancer networks (Olive et al., 2010), in lung cancers with occasional gene amplification, especially in those with small-cell lung cancer histology (Hayashita et al., 2005), breast cancer (Di Leva et al., 2010; Yu et al., 2010), and myeloma (Lionetti et al., 2009).

In this study we applied statistical tests to investigate the relations of the above three chromosomal features, i.e., CpG island, fragile site, and miRNA cluster with cancer-related miRNAs.

In summary, the following issues are addressed in this paper, (i) set up a platform for identifying miRNA-regulate cancer genes given the fact that target gene information are experimentally verified; (ii) quantify the miRNA and target gene expression profile by computing the correlation coefficient; (iii) validate cancer-related miRNA–mRNA pairs based on disease term text mining; and (iv) investigate the relevance of miRNA-associated chromosomal features with cancer.

2. Materials and methods

2.1. Input data

MiRNA information is adopted from a miRBase (Griffiths-Jones et al., 2006). Since target genes may be involved in non-cancerous processes such as cell differentiation, development and normal biological processes; cancer gene information is required in order to identify cancer-related miRNA–mRNA pairs. In this study cancer genes (both oncogene (OCG) and tumor suppressor genes (TSG)) are retrieved from three different resources, i.e., Memorial Sloan-Kettering Cancer Center (MSKCC, http://www.mskcc.org/mskcc/html/44.cfm), Tumor Associate Gene database (TAG) (Chan, 2006) and National Yang Ming University. A total of 659 OCG and 1023 TSG were collected. It is found that 123 genes play a dual role in cancer, that is, they act as an OCG and a TSG. Transcription factor information is also included, which is provided by the TRANSFAC® database (Matys et al., 2006).

The miRNA target gene information is obtained from the TarBase database. TarBase is a manually curated collection of experimentally tested miRNA target genes. Each experimentally validated target site is extracted from the literature. TarBase includes miRNA target gene records for several species, such as human, mouse, fruit fly and worm.

MiRNA and mRNA gene expression profiles were retrieved from NCI-60 data. NCI-60 is a set of 60 human cancer cell lines derived from diverse tissues. These cell lines include nine tissues' miRNA and mRNA expression information, that is, breast cancer, central neural system (CNS) cancer, colon cancer, leukemia, melanoma, non-small cell lung cancer, ovarian cancer, prostate cancer and renal cancer. Four publicly available datasets of gene expression profiles are selected in this study; including the miRNA expression and the Affymetrix U95(A-E), U133A and U133B mRNA expression datasets. Affymetrix mRNA expression datasets adopt three kinds of normalization methods, that is, GCRMA, MAS5 and RMA. Therefore a total of ten datasets are used, comprised of one miRNA dataset and nine Affymetrix RNA expression datasets with different normalization methods. The NCI-60 website provides a tool, called CellMiner (Shankavaram et al., 2009), to query those microarray datasets.

Cancer-related miRNAs, disease genes, and cancer pathway information were downloaded from the miR2Disease (Jiang et al., 2009), Online Mendelian Inheritance in Man (OMIM), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al.,