**Databases and ontologies**

**Linc2GO: A Human LincRNA Function Annotation Resource Based On ceRNA Hypothesis**

Ke Liu†, Zhangming Yan†, Yuchao Li, Zhirong Sun*

MOE Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, School of Life Sciences, Tsinghua University, Beijing 100084, China

**ABSTRACT**

**Summary:** Large numbers of lincRNA (long intergenic non-coding RNA) have been detected through high-throughput sequencing technology. However, currently we still know very little about their functions. Therefore, a lincRNA function annotation database is needed to facilitate the study in this field. In this article, we present Linc2GO, a web resource which aims to provide comprehensive functional annotations for human lincRNA. MicroRNA-mRNA as well as microRNA-lincRNA interaction data were integrated to generate lincRNA functional annotations based on the "ceRNA hypothesis". To the best of our knowledge, Linc2GO is the first database that makes use of the "ceRNA hypothesis" to predict lincRNA functions.

**Availability:** Freely available at http://www.bioinfo.tsinghua.edu.cn/~liuke/Linc2GO/index.html

**Contact:** sunzhr@mail.tsinghua.edu.cn

1 INTRODUCTION

LincRNA (long intergenic non-coding RNA) are endogenous long non-coding RNA molecules that transcribed from “intergenic” regions of the genome. In recent years, as high-throughput sequencing technology develops, more and more lincRNA have been identified (Andrea Pauli, 2011; Guttman, et al., 2009; Young, et al.). It has been demonstrated that lincRNA play critical roles in regulating multiple important biological processes (Guttman, et al.; Guttman and Rinn; Khalil, et al., 2009; Victoria A. Moran, 2012). However, currently most lincRNA are still poorly studied. Due to the importance of lincRNA, it has become very necessary to develop computational tools to predict lincRNA functions. Previous researchers have proposed a “co-expression-based” method to predict lincRNA functions and got many meaningful results (Guo, et al., 2012; Liao, et al.; Loewer, et al.). That is, if a lincRNA is co-expressed with a protein-coding gene whose function is already known, then the lincRNA is predicted to take similar functions.

In this paper we predict lincRNA functions in a rather different approach and present Linc2GO, a novel lincRNA function annotation database. Instead of using the co-expression information of lincRNA and protein coding-gene, our work is based on the competing endogenous RNA hypothesis (short for "ceRNA hypothesis"): lincRNA can function as microRNA “sponge” to interact directly with microRNA and prevent them from binding to mRNA. In such way, lincRNA regulates gene expression and meanwhile regulates the biological processes in which they get involved (Salmena, et al.; Zhao, et al., 2008). For example, MAML1 and MEF2C are two important transcription factors that activate muscle-specific gene expression. Marcella Cesana et al. showed that a lincRNA linc-MD1 regulates muscle differentiation by interacting with two microRNAs miR-135 and miR-133, which can bind to MAML1 and MEF2C to regulate their expressions (Cesana, et al.). Based on the above fact, it is very natural to infer that if a lincRNA and an mRNA share some microRNAs which can interact with both of them, then they two have similar biological functions.

2 METHODS

We downloaded 3 microRNA-mRNA interaction datasets predicted by 3 different algorithms: TargetScan (Bartel, 2007; Lewis, et al., 2005), miRanda (Betel, et al.) and PITA (Kertesz, et al., 2007). Then we integrated them into one dataset, which is much more accurate (See Supplementary Material). Finally, we got 1218961 microRNA-mRNA interactions.

Human lincRNAs are from “Human lincRNA Catalog” (Cabili, et al.). All lincRNA sequences were downloaded from UCSC Genome Browser. MicroRNA sequences were downloaded from miRBase (Kozomara and Griffiths-Jones). The miRanda software is used to predict microRNA-lincRNA interactions with default parameters. We finally got 2198132 human microRNA-lincRNA interactions.

Having got microRNA-mRNA and microRNA-lincRNA interaction data, we followed the principle and workflow shown in Fig.1 to generate lincRNA functional annotations.

First, for each ceRNA-ceRNA pair (the ceRNA-ceRNA pairs here include lincRNA-mRNA pairs, mRNA-mRNA pairs and lincRNA-lincRNA pairs), hypergeometric distribution was used to measure whether the two ceRNAs significantly share some mi-
microRNAs which can interact with both of them. The P-Value was calculated as:

$$P = 1 - \sum_{t=0}^{L} \binom{L}{t} \left( \frac{N - L}{M - t} \right)$$

where N is the total number of microRNA, M is the number of microRNA which interact with the first ceRNA, L is the number of microRNA that interact with the second ceRNA and x is the number of microRNA that interact with both of them. Only the ceRNA-ceRNA pairs with a small P-Value (FDR <0.05) were kept for further analyze.

Next, we combined the kept ceRNA-ceRNA pairs to generate a “ceRNA network” with network nodes represent ceRNA (either lincRNA or mRNA) and the two ceRNAs presented in the same ceRNA-ceRNA pair were connected by an edge. For each lincRNA in the ceRNA network, we collected all its mRNA neighbors and mapped them to corresponding genes, getting a gene set. GO enrichment analyzes and KEGG enrichment analyzes were then applied to the gene set. The final generated GO terms and KEGG pathway terms were assigned to the lincRNA as its functional annotation.

The Lin2GO has provided a very friendly user interface to access functional annotations (See Fig S1). Three additional text files in csv format that contain the all predicted functional annotations are also downloadable to provide convenient access in computational projects.

**ACKNOWLEDGEMENTS**

This work was supported by 973 projects (No. 2009CB918801, No. 2012CB705200) of China, and National Natural Science Foundation of China (NSFC) fund (No.31171274).

**REFERENCES**


