Computational Biology and Chemistry 50 (2014) 60–67

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

piClust: A density based piRNA clustering algorithm

Inuk Jung[a,b], Jong Chan Park[c], Sun Kim[a,b,c,*]

[a] Interdisciplinary Program in Bioinformatics, Republic of Korea
[b] Bioinformatics Institute, Republic of Korea
[c] Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Accepted 23 December 2013
Available online 23 January 2014

Keywords:
piRNA
PIWI
Clustering
RNA-seq
Genome

ABSTRACT

Piwi-interacting RNAs (piRNAs) are recently discovered, endogenous small non-coding RNAs. piRNAs protect the genome from invasive transposable elements (TE) and sustain integrity of the genome in germ cell lineages. Small RNA-sequencing data can be used to detect piRNA activations in a cell under a specific condition. However, identification of cell specific piRNA activations requires sophisticated computational methods. As of now, there is only one computational method, proTRAC, to locate activated piRNAs from the sequencing data. proTRAC detects piRNA clusters based on a probabilistic analysis with assumption of a uniform distribution. Unfortunately, we were not able to locate activated piRNAs from our proprietary sequencing data in chicken germ cells using proTRAC. With a careful investigation on data sets, we found that a uniform or any statistical distribution for detecting piRNA clusters may not be assumed. Furthermore, small RNA-seq data contains many different types of RNAs which was not carefully taken into account in previous studies. To improve piRNA cluster identification, we developed piClust that uses a density based clustering approach without assumption of any parametric distribution. In previous studies, it is known that piRNAs exhibit a strong tendency of forming piRNA clusters in syntenic regions of the genome. Thus, the density based clustering approach is effective and robust to the existence of non-piRNAs or noise in the data. In experiments with piRNA data from human, mouse, rat and chicken, piClust was able to detect piRNA clusters from total small RNA-seq data from germ cell lines, while proTRAC was not successful. piClust outperformed proTRAC in terms of sensitivity and running time (up to 200 folds). piClust is currently available as a web service at http://epigenomics.snu.ac.kr/piclustweb.

1. Introduction

Small non-coding RNAs, interacting with Argonaute (Ago) family proteins, perform functions related to mRNA degradation, transcriptional repression, heterochromatin formation, and DNA elimination in a nucleotide sequence-specific manner (Kim, 2006). The Ago family proteins are divided into two subfamilies: Ago and P-element-induced wimpy testis (PIWI). Ago proteins bind with miRNAs and siRNAs that are expressed in all tissues and regulate mRNAs in eukaryotes. In contrast, expression of PIWI proteins is largely restricted to germ cells and stem cells (Thomson and Lin, 2009). PIWI proteins are known to function as a defensive mechanism against invasive transposable elements (TE) (Kalmykova et al., 2005), which is essential to maintain the integrity of the genome in germ cell lineages. First discovered in 2006, Piwi-interacting RNAs (piRNAs) are novel endogenous non-coding small RNAs that associate with the Piwi protein family. Uncontrolled transposon expression due to the deficiency of Piwi family proteins showed spermatogenesis failure and sterility in mouse (De Fazio et al., 2011) whose descendants also showed to be predisposed to inheriting mutations (Ishizu et al., 2012). Similar results were also observed in porcine testes in the analysis using small RNA-seq data suggesting that piRNAs have role in regulating spermatogenesis (Liu et al., 2012). Another study demonstrated that piRNAs play a role in LINE1 suppression in human HELA cancer cell line (Lu et al., 2010) thus suggesting piRNA activity outside the germ cells.

Several studies show the important roles of PIWI proteins and piRNA. piRNAs mainly arise from intergenic repetitive elements and most of piRNAs have an antisense orientation to active transposon transcripts which is the mechanism of TE silencing. In the ping-pong model (Brennecke et al., 2007), transcripts of transposable elements are cleaved by piRNA-RISC involving Piwi domain proteins such as Aub and Piwi in fly and MILI and MIWI in mouse. This ping-pong cycle simultaneously produces primary piRNAs, processed from a piRNA cluster transcript, and secondary piRNAs, cleaved from an active transposon transcript due to interaction with the piRNA-RISC, resulting in a self amplification cycle for production of piRNA transcripts. Fig. 1 illustrates the ping-pong model.

* Corresponding author at: Seoul National University, Daehak-dong, Gwanak-gu, Seoul, Republic of Korea. Tel.: +82 28802859; fax: +82 28889623.
E-mail addresses: inukjung@snu.ac.kr (I. Jung), uesima@snu.ac.kr (J.C. Park), sunkim.bioinfo@snu.ac.kr (S. Kim).

1476-9271/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2014.01.008

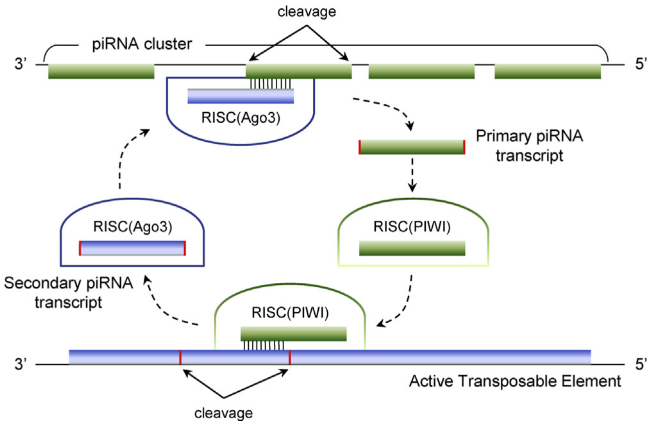

Fig. 1. The ping-pong model.

In general, piRNA sequences have characteristics in terms of length of 25–31 nucleotides (nt) and nucleotide bias at positions 1 and 10 of the 5′ UTR termini. Mammalian piRNAs can be divided into two differentially regulated subclasses referred to pachytene (29–31 nt) and pre-pachytene (26–28 nt) piRNAs with abundance ratio of 10–1 respectively (Aravin et al., 2007). piRNAs are abundant in most metazoa while absent in plants and fungi. A majority of piRNAs (e.g., >83% in mouse) are also known to have unique mapping sites in the genome. Distinct from miRNA and siRNA, piRNAs are thought to be produced through a dicer-independent biogenesis pathway and do not convey sequence conservation across species. However, piRNAs tend to appear in clusters in syntenic regions on the genome and clusters are known to appear across species (Assis and Kondrashov, 2009). Thus, detecting piRNA clusters is a fundamental task for identifying the piRNA transcripts.

There are several studies in regard to the discovery of piRNA and observation of cluster-like distribution of piRNAs in the genome. Studies using piRNAs from mouse testes observed a strong tendency of forming piRNA clusters in syntenic regions of the genome (Girard et al., 2006; Aravin et al., 2006; Grivna et al., 2006; Watanabe et al., 2006; Lau et al., 2006). Another study (Girard et al., 2006) also found similar results in rat and human species. Until now, there is only one database, piRNABank (Sai Lakshmi and Agrawal, 2007), that provides a collection of piRNAs from human, mouse, rat, and Macaca mulatta from aforementioned studies and 454-NGS sequencing in the course of the development of the proTRAC software (Rosenkranz and Zischler, 2012).

1.1. The piRNA detection problem

Our package, piClust, takes a small RNA-seq data as input and produces as output a set of piRNAs that are position-wise clustered on the genome, i.e., piRNA clusters. Before computing piRNA clusters, all known non-coding RNAs are discarded from the input small RNA-seq data. Then the remaining RNA-seq data is mapped to the reference genome, which is input to the clustering step. The remaining RNA-seq data still contains non-piRNAs, thus the computational challenge is to distinguish piRNAs from non-piRNAs given the alignment file where short reads are mapped to the reference genome. Since it is known that piRNAs are known to cluster position-wise on the genome, we use a density based clustering approach to predict piRNAs in the alignment file. However, it is misleading to assume that all such reads in clusters are real piRNAs. Thus, we also use known piRNA characteristics to further remove non-piRNAs in candidate piRNA clusters.

1.2. Motivation

There is only one software, proTRAC, for detecting piRNA clusters. proTRAC is based on a statistical probabilistic analysis. Assuming a uniform distribution, proTRAC determines clusters based on the number of aligned putative piRNAs within a 1 kbp (1000 bp) sliding window. If the number of putative piRNA within a 1 kbp window is significantly high (i.e., p-value <0.01), proTRAC begins to track the region as a cluster. Each detected cluster is then further examined by a probabilistic scoring scheme considering whether the cluster meets piRNA cluster characteristics. Another recent study, piRNA predictor (Zhang et al., 2011), presented a method to predict piRNA transcripts rather than predicting piRNA clusters using a k-mer method. This method can be seen as a filtering method based on k-mers that can be applied to piRNA transcripts after piRNA clusters are detected.

1.2.1. Weakness of current piRNA cluster detection approaches

1.2.1.1. Use of arbitrary window size for observing density. proTRAC uses a sliding window approach for detecting piRNA clusters. The window size is a cluster parameter that should be set reflecting the characteristics of the data set. The length of currently annotated piRNA clusters in human, mouse and rats varies ranging widely from 1 kbp to 100 kbp. Thus, current piRNA cluster detecting approaches that use an arbitrary window size without clear definition may not be effective.

1.2.1.2. Assumption of statistical distributions. proTRAC defines piRNA clusters when they significantly deviate from a uniform distribution assuming that non-piRNAs follow the uniform distribution. However, with a careful investigation on data sets, we found that a uniform or any statistical distribution for non-piRNAs may not be assumed. To test for uniform distribution, we counted the number of alignments of putatively non-piRNA transcripts within a 1 kbp sliding window for each chromosome and performed a Chi-square test. The Chi-square test returned p-values lower than $2.2e^{-6}$ for all chromosomes which suggests that non-piRNA does not follow a uniform distribution. Furthermore, testing for normality was also rejected through a Q–Q plot. Hence, use of a statistical distribution may not be suitable for detecting piRNA clusters.

1.2.1.3. Misleading definition of piRNA for calculating clustering parameters. proTRAC statistically determines the minimum density for a region to be tracked as a cluster based on the binomial distribution. proTRAC considers every transcript within a certain length range (e.g., 25–31 nt), and examines whether the transcripts in the cluster meet the minimum density using the following equation:

$$p(n \geq k) = \Sigma_{k=n}^{1000} \binom{1000}{k} r^k (1-r)^{1000-k}$$

where k is the expected number of alignments in a 1000 bp window and r is the ratio of transcripts with proper piRNA length within the entire data set. The minimum number of loci per 1 kbp is set to n when $p(n \geq k) = 0.01$.

However, it is difficult to make such judgment with total small RNA-seq data that contains all kinds of small RNAs, including piRNA. Since some transcripts may not actually associate with PIWI proteins, considering every transcript with proper piRNA length as a real piRNA transcript can be misleading. Even with piRNA specifically curated data, Girard et al. discarded reads that have more than five mapping sites to avoid false positives.