BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 22 2011, pages 3202–3203
doi:10.1093/bioinformatics/btr527

Sequence analysis

Advance Access publication October 5, 2011

A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome

Lorena Pantano[1], Xavier Estivill[1,2,*] and Eulalia Martí[1,2,*]

[1]Genetic Causes of Disease, Genes and Disease Programme, Centre for Genomic Regulation (CRG) and UPF and [2]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Recent progress in high-throughput sequencing technologies has largely contributed to reveal a highly complex landscape of small non-coding RNAs (sRNAs), including novel non-canonical sRNAs derived from long non-coding RNA, repeated elements, transcription start sites and splicing site regions among others. The published frameworks for sRNA data analysis are focused on miRNA detection and prediction, ignoring further information in the dataset. As a consequence, tools for the identification and classification of the sRNAs not belonging to miRNA family are currently lacking.

Results: Here, we present, SeqCluster, an extension of the currently available SeqBuster tool to identify and analyze at different levels the sRNAs not annotated or predicted as miRNAs. This new module deals with sequences mapping onto multiple locations and permits a highly versatile and user-friendly interaction with the data in order to easily classify sRNA sequences with a putative functional importance. We were able to detect all known classes of sRNAs described to date using SeqCluster with different sRNA datasets.

Availability: tool and video-tutorials are available at http://estivill_lab.crg.es/seqbuster.

Contact: eulalia.marti@crg.es; xavier.estivill@crg.es

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 14, 2011; revised on September 13, 2011; accepted on September 15, 2011

1 INTRODUCTION

Small silencing RNAs are the best-known class of non-coding RNAs (ncRNAs) of 18–30 nt in length, involved in gene silencing by association to the Argonaute family of proteins (Czech and Hannon, 2011). Recent progress in high-throughput sequencing technologies has largely contributed to elucidate the remarkable landscape of sRNAs, revealing new species of sRNAs with unknown functions. These novel sRNAs are classified according to their position in the genome and putative functions. Nowadays, five major groups of sRNAs have been detected: (i) tiRNAs, located at the transcription initiation site of coding genes (Taft *et al.*, 2010); (ii) spli-RNAs, detected at splicing site of transcripts (Taft *et al.*, 2010); (iii) tasi-RNAs, associated to gene termini; (iv) sRNAs derived from tRNA (Haussecker *et al.*, 2010) and (v) sRNAs from non-coding RNA regions (Ono *et al.*, 2011). No tools are prepared to cover a complete analysis of data coming from sRNA sequencing, and the existing ones are only for miRNA characterization and prediction. A major challenging problem using high-throughput sequencing data is annotation when the sequences map onto multiple locations. The current frameworks resolve this situation with heuristic assumptions, including non-consistent data removal or providing random annotations. This produces biased results that hamper the discovery and classification of novel sRNAs. Here, we present SeqCluster, a tool for the characterization of the non-miRNA sRNA transcriptome. SeqCluster is presented as an extension of SeqBuster, a pipeline for the characterization of miRNAs (Pantano *et al.*, 2010) and constitutes the first framework giving a completely unbiased classification of non-miRNAs data of any species. SeqCluster permits a user-friendly interaction with the data at any level in order to easily classify and annotate small RNA sequences with a putative functional importance.

2 METHODS

SeqCluster, an extension of the miRNA-analysis tool SeqBuster, has been developed to analyze any kind of sRNA detected by large-scale sequencing technologies (see 'implementation' in Supplementary Material). The new framework integrates three specific processes: (1) raw data processing and miRNA detection, (2) clustering and (3) classification (Fig. 1). In the first process, the adapter is trimmed from the raw sequences, and subsequently sequences are mapped onto miRNA and miRNA precursor databases. To avoid the dependency on an external tool for mapping against miRBase dataset (Griffiths-Jones, 2004), a custom algorithm based on seed (fragments of 8 nt) indexation has been integrated in Java (see 'miRNA detection' in Supplementary Material). miRNAs predicted using an external tool may be loaded to SeqCluster to avoid the incorporation of these sequences to the study of the non-miRNA sRNA transcriptome. The second step defines unit-sRNAs (usRNAs) taking into account two filters: (i) sequence similarity and (ii) genome location. In the first filter, all sequences with 100% identity (no mismatches allowed) and >80% of overlapping are considered as putative unit small RNA (pre-usRNA). In the second filter, all pre-usRNAs are mapped onto a custom genome using megablast (Altschul *et al.*, 1990) or other custom tool. Otherwise, annotated data from any other mapping tool may be directly uploaded onto SeqCluster. This filter only affects ambiguously overlapped pre-usRNA and are used to make the decision on whether or not the two overlapping pre-usRNAs should be considered a single cluster of sequences (usRNA). In the rest of the cases (unique sequences or unambiguously mapped clustered sequences), pre-usRNAs are directly called as usRNAs. The requirement to merge ambiguous overlapped pre-usRNA into ambiguous usRNA is that all overlapped pre-usRNAs have to share all the same regions. When the latter does not occur due to more complex situations, pre-usRNAs enter into an extra module integrating a recursive algorithm (see 'decision algorithm' in Supplementary Material).

*To whom correspondence should be addressed.

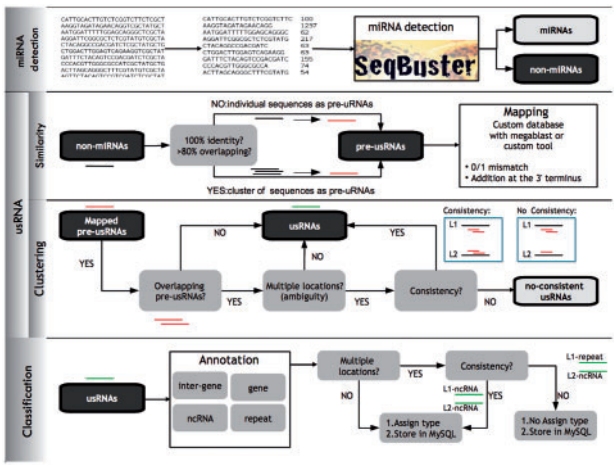

Fig. 1. SeqCluster extension framework scheme. The framework integrates three specific processes: (1) raw data processing and miRNA detection, (2) clustering and (3) classification. In the first process the adapter is trimmed from the raw sequences, and subsequently sequences are mapped onto miRNA and miRNA precursor databases. After that, all sequences not annotated as miRNAs or miRNA precursors are clustered in clustering step. This step is performed according to two filters: (i) sequence similarity and (ii) genome location. In the classification step, usRNAs are classified according to the genome context that will cover as many classes and subclasses as the users define. Once usRNAs are defined and located on the genome, all of them are classified according to the genome context that will cover as many classes as the users define, being the more common types: non-coding RNAs, transposable elements (TEs), and genes (see 'usRNA' classification in Supplementary Material).

3 OUTPUT

The framework generates a main MySQL table for each sample where rows are usRNAs and columns show the following information: unique identifier, number of sequences, number of locations, coordinates and finally one column for each class, according to the annotation step: -repeat, -ncRNA, etc (see 'output scheme' in Supplementary Material). For a user-friendly view, BED files are generated to be uploaded to UCSC (Kent *et al.*, 2002). Furthermore, SeqCluster permits differential expression analysis between two samples or two groups of samples in different biological contexts. Datasets involving time series experiments may be also analyzed .

4 RESULTS

We have applied SeqCluster extension to analyze small RNA datasets of human brain samples sequenced by illumina 1G in our previous work (Martí *et al.*, 2010) and other public dataset from different species (see 'SeqCluster application' to real datasets in Supplementary Material). First, we detected miRNA sequences using SeqBuster with default parameters and the miRBase resource (Release 15, (Griffiths-Jones, 2004)). We applied SeqCluster extension to the rest of the data, resulting in a total of 8335, 12 366, 15 614, 44 265 and 82 985 sequences annotated as usRNAs in human brain, human stem cells, mouse, fly and worm, respectively.

5 CONCLUSIONS

Differing from the current sRNA analysis tools, the main advantages of SeqCluster framework are as follows: (i) the classification and annotation of the data are not restricted to specific databases, offering the possibility to perform this analysis with any custom database; (ii) the implementation of filters integrated to solve, in a non-biased way, the problem of ambiguous sRNAs mapping; (iii) the opportunity to further study the presumed biogenesis of the sRNAs and; (iv) the possibility to inspect highly expressed sequences that have not been successfully classified allowing the extraction of complex sRNAs for further analysis. Our results validate SeqCluster as a tool to detect and classify all types of sRNAs in different species, including the most recently discovered classes of still unknown function (Taft *et al.*, 2010).

Funding: Spanish Ministry of Health Fondo de Investigaciones Sanitarias (PI081367) and Instituto de Salud Carlos III (CIBERESP); Spanish Ministry of Science and Innovation (SAF2008-00357) the Sixth Framework Programme of the European Commission through the SIROCCO integrated project LSHG-CT-2006-037900 and the Spanish Ministry of Science and Innovation (SAF2008-00357). E.M. is partially supported by the Spanish Ministry of Health; L.P. is recipient of a fellowship from the Spanish Ministry of Science and Innovation MICINN.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Czech,B. and Hannon,G.J. (2011) Small rna sorting: matchmaking for argonautes. *Nat. Rev. Genet.*, **12**, 19–31.

Griffiths-Jones,S. (2004) The microrna registry. *Nucleic Acids Res.*, **32**, 109–111.

Haussecker,D. *et al.* (2010) Human trna-derived small rnas in the global regulation of rna silencing. *RNA*, **16**, 673–695.

Kent,W.J. *et al.* (2002) The human genome browser at ucsc. *Genome Res.*, **12**, 996–1006.

Martí,E. *et al.* (2010) A myriad of mirna variants in control and huntington's disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res.*, **38**, 7219–7235.

Ono,M. *et al.* (2011) Identification of human mirna precursors that resemble box c/d snornas. *Nucleic Acids Res.*, **39**, 3879–3891.

Pantano,L. *et al.* (2010) Seqbuster, a bioinformatic tool for the processing and analysis of small rnas datasets, reveals ubiquitous mirna modifications in human embryonic cells. *Nucleic Acids Res.*, **38**, e34.

Taft,R.J. *et al.* (2010) Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. *Nat. Struct. Mol. Biol.*, **17**, 1030–1034.