BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 14 2012, pages 1925–1927
doi:10.1093/bioinformatics/bts282

Sequence analysis Advance Access publication May 9, 2012

CPSS: a computational platform for the analysis of small RNA deep sequencing data

Yuanwei Zhang[1,†], Bo Xu[1,†], Yifan Yang[2], Rongjun Ban[3], Huan Zhang[1], Xiaohua Jiang[1], Howard J. Cooke[1,4], Yu Xue[5,*] and Qinghua Shi[1,*]

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei 230027, China, [2]Department of Statistics, University of Kentucky, Lexington, KY 40506, USA, [3]Department of Computer Science & Technology, Nanjing University, Nanjing 210093, [4]MRC Human Genetics Unit, IGMM, University of Edinburgh, Edinburgh EH4 2XU, UK, and [5]Department of Biomedical Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: Next generation sequencing (NGS) techniques have been widely used to document the small ribonucleic acids (RNAs) implicated in a variety of biological, physiological and pathological processes. An integrated computational tool is needed for handling and analysing the enormous datasets from small RNA deep sequencing approach. Herein, we present a novel web server, CPSS (a computational platform for the analysis of small RNA deep sequencing data), designed to completely annotate and functionally analyse microRNAs (miRNAs) from NGS data on one platform with a single data submission. Small RNA NGS data can be submitted to this server with analysis results being returned in two parts: (i) annotation analysis, which provides the most comprehensive analysis for small RNA transcriptome, including length distribution and genome mapping of sequencing reads, small RNA quantification, prediction of novel miRNAs, identification of differentially expressed miRNAs, piwi-interacting RNAs and other non-coding small RNAs between paired samples and detection of miRNA editing and modifications and (ii) functional analysis, including prediction of miRNA targeted genes by multiple tools, enrichment of gene ontology terms, signalling pathway involvement and protein–protein interaction analysis for the predicted genes. CPSS, a ready-to-use web server that integrates most functions of currently available bioinformatics tools, provides all the information wanted by the majority of users from small RNA deep sequencing datasets.

Availability: CPSS is implemented in PHP/PERL+MySQL+R and can be freely accessed at http://mcg.ustc.edu.cn/db/cpss/index.html or http://mcg.ustc.edu.cn/sdap1/cpss/index.html.

Contact: xueyu@mail.hust.edu.cn or qshi@ustc.edu.cn

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

Received on August 25, 2011; revised on April 20, 2012; accepted on May 3, 2012

1 INTRODUCTION

Non-coding ribonucleic acids (RNAs), which do not encode proteins, include ribosomal RNAs, transfer RNAs, microRNAs (miRNA), piwi-interacting RNAs (piRNAs) and other RNA species. These RNAs participate in a surprisingly diverse collection of regulatory events (Moazed, 2009). miRNAs have received particular attention due to their negative role in widespread regulation of mRNA metabolism through direct base pairing interactions at transcriptional and post-transcriptional levels (Carthew and Sontheimer, 2009). To better understand the regulatory roles of miRNAs and other small RNAs in different tissues and developmental stages, the expression profiles of small RNAs need to be assessed.

Recently, the emergence of the next generation sequencing (NGS) techniques has revolutionized the identification of small RNAs with particularly high levels of sensitivity and accuracy (Zhou *et al.*, 2011). Several published tools detecting non-coding RNA profiles from NGS data have been developed. For example, miRExpress (Wang *et al.*, 2009) is a stand-alone software for detecting known miRNAs and novel miRNAs. miRanalyzer (Hackenberg *et al.*, 2009), which also offers stand-alone version, is a web server tool that can detect known and novel miRNAs, identify differentially expressed miRNAs and predict miRNA targets. SeqBuster (Pantano *et al.*, 2010), offering a web-based toolkit and stand-alone version, focuses on detecting miRNA variants/isoforms for known miRNAs and can also be used to identify differentially expressed miRNAs and predict miRNA targets. There are several recent comprehensive tools designed to analyse NGS data. mirTools (Zhu *et al.*, 2010) is a web-based tool designed to explore the genome map and length distribution of short reads and to classify them into known categories, to detect differentially expressed miRNAs and to predict novel miRNAs and their secondary structures. DARIO (Fasold *et al.*, 2011) is a free web service for detecting and normalizing non-coding RNA expression and predicting novel non-coding RNAs. WapRNA (Zhao *et al.*, 2011) is not used only to detect miRNA expression profile from small RNA NGS data but also to analyse mRNA NGS data (detailed overview of CPSS and other 13 bioinformtics tools are shown in Supplementary Table S1). Until now, none of the currently available tools provides functional analysis for predicted targets of miRNAs from NGS data, which could help users to find potentially

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

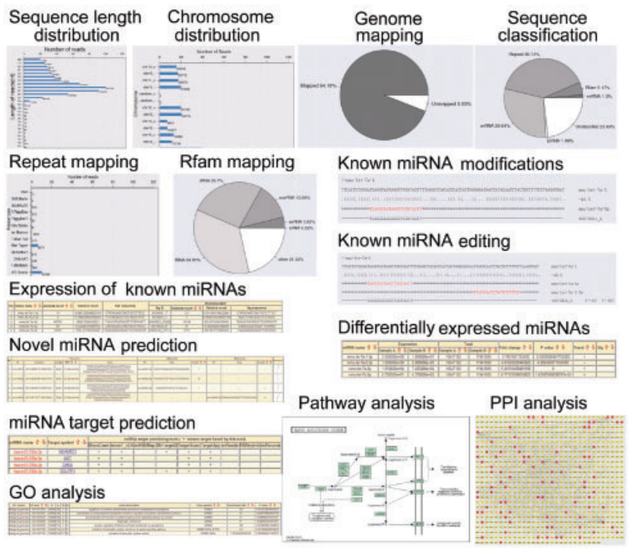

Fig. 1. The summary results of CPSS (details are shown in http://mcg.ustc.edu.cn/db/cpss/download/6818289109/result_index.html)

candidate genes/pathways for further experimental or computational studies. Thus, a comprehensive and systematic tool, integrating most features of previous tools with functional analysis for predicted targets of miRNAs from NGS data, is still needed.

Herein, we present a novel and free web server, CPSS, which integrates most functions of currently available bioinformatics tools (Supplementary Table S1 and Supplementary Fig. S1). By using CPSS, small RNA NGS data can be analysed systematically in one platform after a single submission of data by integration of annotation and functional analysis of novel and/or differentially expressed miRNAs. CPSS generates an analysis report including: (i) annotation analysis, which provides a comprehensive analysis for small RNA transcriptome, such as length distribution and genome mapping of sequencing reads, small RNA annotation, prediction of novel miRNAs, identification of differentially expressed miRNAs, piRNAs and other non-coding small RNAs between paired samples and detection of miRNA editing and modifications and (ii) functional analysis, which provides the functional analysis of miRNAs, e.g. predicting miRNA target genes by multi-tools, enriching gene ontology (GO) terms, performing signalling pathways and analysing protein–protein interaction (PPI) for the predicted genes (Fig. 1).

2 WORKFLOW

CPSS provides an easy-to-use interface, allowing users to conveniently analyse the data of small RNA transcriptome from NGS techniques. Users submit the input data in FASTA format or FASTA files compressed in *.gz format, and the FASTA format (*.fa) files can be transformed from the raw data, which contain unprecedented amounts of reads generated by Illumina Genome Analyzer, 454 FLX instrument or SOLiD^{TM} system. The overall workflow of CPSS is shown in Supplementary Figure S2. First, the remaining clean sequences in FASTA format filtered above are classified into several categories, i.e. miRNA, piRNA, other non-coding small RNAs, mRNA, genomic repeats, etc. Then, the sequences that can be mapped to the reference genome but cannot be assigned to any of the referred annotations are used to predict novel miRNAs using mireap (http://sourceforge.net/projects/mireap) or miRDeep (Friedlander *et al.*, 2008), and their secondary structures are predicted by RNAfold (http://rna.tbi.univie.ac.at/). The potential target genes will be predicted for the most abundant novel/known miRNAs from one sample and for all the differentially expressed miRNAs from two paired samples automatically by eight miRNA target prediction tools. For functional analysis of miRNAs, the predicted targets are mapped to the GO annotation dataset (Ashburner *et al.*, 2000) and used to extract the enriched GO processes using the Fisher's exact test (enrichment ratio >2 and $P<0.01$). Then, the genes in the enriched GO processes are matched to the signal pathway annotation datasets from KEGG (Kanehisa *et al.*, 2010) and PPI annotation dataset from String (Jensen *et al.*, 2009). (The details of algorithm for every step are presented in Supplementary Material.) CPSS is ready-to-use for most users without the need to change any of the default analysis parameters. However, users can also modify most of the parameters according to their advanced requirements. The final results are presented to users as graphic summary in a browser (Fig. 1) and detailed results are saved in a *.gz file that can be downloaded from the server.

Currently, CPSS is able to handle the data from either one or two samples on a centralized platform and can complete a job within 0.5–3 h, depending on data size, species and selection of parameters. This server displays the status of a job and sends a reminder email to users when the job is done. Users can retrieve the analysis results from the stored jobs with a unique ID generated randomly by the server for each job. Most strikingly, the annotation and functional analysis of novel and/or differentially expressed miRNAs from small RNA NGS data can be completed in one platform, CPSS, after a single submission of data.

2.1 Case studies

To evaluate the performance of CPSS, several small RNA sequencing data from our laboratory are tested. First, one sample from human ovary was uploaded to CPSS (Zhang *et al.*, 2011). Standard protocols were used for small RNA preparation and Illumina sequencing. In total, 8 721 844 clean reads were generated for the sample. According to the workflow, we filtered and annotated them using currently available databases (Supplementary Table S2). Massively parallel pyrosequencing generated 11 966 289 non-redundant sequences from the human ovary, and 260 predicted novel miRNAs were detected (Supplementary Fig. S3A). All the annotation and analysis were completed in 30 min and the detailed results are available from CPSS (Download ID = 6818289109). Second, two samples of small RNA sequencing data from testes of Spo11 knockout and wild-type mice were tested. Following the workflow, all the small RNA sequences were also filtered and annotated using CPSS based on the currently available databases, and the significant differentially expressed miRNAs and piRNAs between the samples were detected and functional analysis for them was completed (Supplementary Fig. S3B) within 1 h (Download ID = 713628095). The differential expression of these miRNAs and piRNAs (these miRNAs and piRNAs expressed differentially based on both total read counts and most abundant unique tag) was further validated by real-time polymerase chain reaction (PCR) (details in Supplementary Methods), and a strong correlation for miRNA levels was detected between deep sequencing and real-time PCR

(Supplementary Fig. S3C and D; $R = 0.947$ and 0.916, respectively), indicating the credibility and robustness of deep sequencing-based expression analysis obtained from CPSS.

Funding: National Basic Research Program [2012CB944402, 2007CB947401, 2011CB944501] of China (973); Program of Knowledge Innovation [KSCX2-EW-R-07] of Chinese Academy of Science, China.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.

Carthew,R.W. and Sontheimer,E.J. (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell*, **136**, 642–655.

Fasold,M. *et al.* (2011) DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res*., **39**, 112–117.

Friedlander,M.R. *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol*., **26**, 407–415.

Hackenberg,M. *et al*. (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, **37**, W68–W76.

Jensen,L.J. *et al*. (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.

Kanehisa,M. *et al*. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Moazed,D. (2009) Small RNAs in transcriptional gene silencing and genome defence. *Nature*, **457**, 413–420.

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.

Wang,W.C. *et al.* (2009) miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics*, **10**, 328.

Zhang,Y.W. *et al.* (2011) Prediction of novel pre-microRNAs with high accuracy through boosting and SVM. *Bioinformatics*, **27**, 1436–1437.

Zhao,W. *et al.* (2011) wapRNA: a web-based application for the processing of RNA sequences. *Bioinformatics*, **27**, 3076–3077.

Zhou,L. *et al.* (2011) Small RNA transcriptome investigation based on next-generation sequencing technology. *J. Genet. Genomics*, **38**, 505–513.

Zhu,E. *et al*. (2010) mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res.*, **38**, W392–W397.