BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 21 2011, pages 3076–3077
doi:10.1093/bioinformatics/btr504

Databases and ontologies Advance Access publication September 6, 2011

wapRNA: a web-based application for the processing of RNA sequences

Wenming Zhao[1,†], Wanfei Liu[1,2,†], Dongmei Tian[1,†], Bixia Tang[1,†], Yanqing Wang[1], Caixia Yu[1], Rujiao Li[1], Yunchao Ling[1,2], Jiayan Wu[1], Shuhui Song[1,*] and Songnian Hu[1,*]

[1]Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029 and [2]Graduate University of Chinese Academy of Sciences, Beijing 100049, P. R. China

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: mRNA/miRNA-seq technology is becoming the leading technology to globally profile gene expression and elucidate the transcriptional regulation mechanisms in living cells. Although there are many tools available for analyzing RNA-seq data, few of them are available as easy accessible online web tools for processing both mRNA and miRNA data for the RNA-seq based user community. As such, we have developed a comprehensive web application tool for processing mRNA-seq and miRNA-seq data. Our web tool wapRNA includes four different modules: mRNA-seq and miRNA-seq sequenced from SOLiD or Solexa platform and all the modules were tested on previously published experimental data. We accept raw sequence data with an optional reads filter, followed by mapping and gene annotation or miRNA prediction. wapRNA also integrates downstream functional analyses such as Gene Ontology, KEGG pathway, miRNA targets prediction and comparison of gene's or miRNA's different expression in different samples. Moreover, we provide the executable packages for installation on user's local server.

Availability: wapRNA is freely available for use at http://waprna.big.ac.cn. The executable packages and the instruction for installation can be downloaded from our web site.

Contact: husn@big.ac.cn; songshh@big.ac.cn

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

Received on May 6, 2011; revised on August 26, 2011; accepted on August 31, 2011

1 INTRODUCTION

RNA-Seq is a recently developed revolutionary approach for transcriptome profiling and miRNA analysis using next generation sequencing technologies. It can allow us to study RNA in further depth and more accurate measurement. Studies using this method have already altered our view of transcriptome and miRNA. Recently, a lot of methods have been developed to process the different aspects of RNA-seq data, such as short reads mapping (Li and Durbin, 2009), gene annotation and differential gene-expression analysis (Wang *et al.*, 2010). Meanwhile, several web-services for RNA and microRNA analysis have also been developed, such as rQuant.web (Bohnert and Ratsch, 2010), MAGIA (Sales *et al.*, 2010), miRCat (Moxon *et al.*, 2008), miRAnalyzer (Hackenberg *et al.*, 2009), mirTools (Zhu *et al.*, 2010) and DSAP (Huang *et al.*, 2010). However, some of these web services need prior data processing, such as mapping raw sequences, normalizing expression data, or they do not provide local installation package.

In order to provide a service for the increasing amount of experimental studies using RNA-seq methods, here, we present the wapRNA, a novel web service that allows us to do mRNA and miRNA expression analysis simultaneously. wapRNA complements other RNA analysis tools reported by Goncalves *et al.* (2011). Moreover, we provide the executable packages for installation on user's local server. This will facilitate studies for RNA using high-throughput approaches, especially for biologists who have limited computing sources or are short of bioinformatics processing experience.

2 METHODS

2.1 The analysis pipeline

The wapRNA web service simplifies the analysis of mRNA-seq/miRNA-seq data by executing a computational pipeline on a set of data uploaded by user. To accommodate different sequencing methods (SOLiD or Solexa) and study targets (RNA or miRNA); we built four modules for SOLiD mRNA, Solexa mRNA, SOLiD miRNA and Solexa miRNA data analysis, respectively. Ten animal species whose genome sequences are available are provided in our server (The species will be added according to user's requirements in future), and the gene structure information is based on ENSEMBL (http://asia.ensembl.org/index.html) annotation. For mRNA data analysis, the process starts from selecting sequencing platform, uploading raw data and specifying the parameters like species and mapping mismatches allowance. When submitted the above request, wapRNA will launch the processes by preparing reference sequences and filtering raw data according to user's request or default parameters. Then, those cleaned reads will be mapped to corresponding reference sequences automatically using Corona_Lite package for SOLiD (http://solidsoftwaretools.com/gf/project/corona/) and BWA package (Li and Durbin, 2009) for Solexa, respectively. The expression levels (RPKM, Reads Per Kilobase of exon model per Million mapped reads) for each gene will be calculated using uniquely mapped reads by in-house built Perl module. Moreover, the downstream functional classification will be launched through the integrated localization of Gene Ontology (Harris *et al.*, 2004) and KEGG pathway database (Kanehisa and Goto, 2000). When the job is finished, an email alerts will be automatically sent if the email address is provided, and all results can be found and downloaded at the results page, and also some concise pictures and statistical tables are provided for users.

While for miRNA data analysis, users also should specify the requisite parameters listed in webpage and submit the job. We adopted two different methods for SOLiD and Solexa data analysis. For Solexa platform,

*To whom correspondence should be addressed.

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

we first remove adaptors sequences, cluster those identical reads into unique tags, map tags to assigned genome sequences and use those perfect hits for the following analysis. Then we filter those non-miRNA (including snRNA, tRNA, rRNA, mRNA, etc.) tags and annotate conserved miRNA by comparing tags to miRBase (Griffiths-Jones, 2006; Griffiths-Jones *et al.*, 2008) database. Finally, we predicate new miRNAs using miRdeep software (Friedlander *et al.*, 2008). While for SOLiD platform, due to the color-spaced sequences, we integrate the RNA2MAP package (http://solidsoftwaretools.com/gf/project/rna2map/) for filtering the non-miRNA sequences, annotating the known miRNAs by aligning to the miRNA precursor sequences from miRBase, and mapping to the genome sequence for predicting novel miRNAs. The miRNA targets are predicated and combined by using RNAhybrid (Kruger and Rehmsmeier, 2006) and miRanda (Betel *et al.*, 2010) software.

Moreover, we developed an independent module for detecting differentially expressed genes (DEG), functional classification (based on GO and KEGG database), and miRNA target predication.

2.2 Case study

To demonstrate the functionality of wapRNA, we used it to analyze SOLiD mRNA data from SRA009022 (Cui *et al.*, 2010) and miRNA data (Cai *et al.*, 2010) produced in our lab. We also tested our Solexa pipeline using mRNA data from SRR037165 (Wu *et al.*, 2010) and Solexa miRNA data from SRR026761 (Morin *et al.*, 2008). All analyzed results were showed in demo results part, and all these jobs were run on a Red Hat Enterprise Linux Cluster with 12 16 GB-RAM compute nodes and a total of 96 CPU cores.

For the SOLiD mRNA pipeline, we got 13.6% multiple mapped reads and 32.8% unique mapped reads, which are relatively higher than Cui *et al.*'s (2010) original results. The difference is caused by parameter setting in mapping procedure. For the miRNA pipeline, we identified 397 (341 unique) known miRNA and 138 (111) novel miRNA in only one sample (hESC), while 315 known miRNA and 100 novel miRNA in hESC samples by Morin *et al.*'s (2008) results, about 82% of those known miRNA and 87% of those novel miRNAs are overlapped between the two methods. The other two tests also showed very similar results with original results.

3 IMPLEMENTATION

wapRNA consists of front- and back-end programs. The front-end program, constructed based on Java Server Page (JSP) 2.0 technology, is in charge of tasks submission and results display. We used the Struts (http://struts.apache.org), Spring (http://www.springsource.org) and Hibernate (http://www.hibernate.org/) framework to enhance the flexibility and extendibility, and Mysql5.0 as the DBMS to store the interim and final results. The back-end program consists of series of Perl and Shell scripts, it is mainly in charge of the data processing.

The executable packages for four pipelines designed and implemented using Java program, which can be download from our web site. All the packages can be run on 64-bit Linux system. Before running the packages, some essential tools declared in the download page are needed to be pre-installed in the system.

4 DISCUSSION

The wapRNA web service provides users to analyze next-generation sequencing mRNA and miRNA data directly and simply by an easy-to-use browser user interface. The service includes major processes for the next-generation mRNA or miRNA data analysis, including preprocessing raw sequenced reads, mapping tags to reference sequences, gene expression annotation, and other downstream functional analysis such as detecting differentially expressed genes, Gene Ontology and KEGG pathway analysis for RNA, novel miRNA predication and miRNA target prediction. Moreover, we also provide executable packages for users to build their pipeline locally. However, some steps are still improvable, for example, we ignored those multiple mapping reads which may be very important for duplicated genes or members of gene family. And also, there are a few additional features we have not involved, such as new transcript identification, alternative splicing, these features will be integrated into the wapRNA in the near future.

ACKNOWLEDGEMENTS

We would like to thank all the members of Genome and Bioinformatics platform in BIG for their assistance during the pipeline development, and thank Peng Cui, Xiaomin Yu for the previous work on the RNA and miRNA research, and also thank Kesheng Liu for the solid data support.

Funding: Natural Science Foundation of China (30900831 to S.S.); Ministry of Science and Technology of the People's Republic of China (BSDN2009-15 to W.Z.); Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-R-01-04 to S.H.).

Conflict of Interest: none declared.

REFERENCES

Betel,D. *et al.* (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.

Bohnert,R. and Ratsch,G. (2010) rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.*, **38**, W348–W351.

Cai,Y. *et al.* (2010) Novel microRNAs in silkworm (Bombyx mori). *Funct. Integr. Genomics*, **10**, 405–415.

Cui,P. *et al.* (2010) A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, **96**, 259–265.

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

Goncalves,A. *et al.* (2011) A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*, **27**, 867–869.

Griffiths-Jones,S. (2006) miRBase: the microRNA sequence database. *Methods Mol. Biol.*, **342**, 129–138.

Griffiths-Jones,S. *et al.* (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.

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

Harris,M.A. *et al.*; Consortium,G.O. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Huang,P.J. *et al.* (2010) DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res.*, **38**, W385–W391.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754–1760.

Morin R.D. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.

Moxon,S. *et al.* (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, **24**, 2252–2253.

Sales,G. *et al.* (2010) MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Res.*, **38**, W352–W359.

Wang,L.K. *et al.* (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.

Wu,J.Q. *et al.* (2010) Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc. Natl Acad. Sci. USA*, **107**, 5254–5259.

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