BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 20 2010, pages 2615–2616
doi:10.1093/bioinformatics/btq493

Genome analysis

Advance Access publication August 27, 2010

miRNAkey: a software for microRNA deep sequencing analysis

Roy Ronen[1], Ido Gan[1], Shira Modai[1], Alona Sukacheov[2], Gideon Dror[2], Eran Halperin[3,4] and Noam Shomron[1,*]

[1]Department of Cell and Developmental Biology, Sackler Faculty of Medicine, Tel Aviv University, [2]The Academic College of Tel-Aviv-Yaffo, Tel-Aviv, Israel, [3]International Computer Science Institute, Berkeley, CA, USA and [4]School of Computer Science and Department of Molecular Microbiology and Biotechnology, George Wise Faculty of Life Science, Tel-Aviv University, Tel-Aviv, Israel

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MicroRNAs (miRNAs) are short abundant non-coding RNAs critical for many cellular processes. Deep sequencing (next-generation sequencing) technologies are being readily used to receive a more accurate depiction of miRNA expression profiles in living cells. This type of analysis is a key step towards improving our understanding of the complexity and mode of miRNA regulation.

Results: miRNAkey is a software package designed to be used as a base-station for the analysis of miRNA deep sequencing data. The package implements common steps taken in the analysis of such data, as well as adds unique features, such as data statistics and multiple read determination, generating a novel platform for the analysis of miRNA expression. A user-friendly graphical interface is applied to determine the analysis steps. The tabular and graphical output contains general and detailed reports on the sequence reads and provides an accurate picture of the differentially expressed miRNAs in paired samples.

Availability and implementation: See http://ibis.tau.ac.il/miRNAkey

Contact: nshomron@post.tau.ac.il

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

Received on July 1, 2010; revised on August 18, 2010; accepted on August 23, 2010

1 INTRODUCTION

MicroRNAs (miRNAs) are short non-coding RNAs abundant in all animal cells. These small non-coding molecules bind to mRNA transcripts and control their expression through facilitated mRNA degradation or protein inhibition (Bushati and Cohen, 2007). miRNAs are involved in many cellular processes suggesting their dominant role in regulation of gene expression (Shomron *et al*., 2009). MiRNA expression profiling, obtained under different conditions or from cells of different origins, provides a basis for the investigation of the regulatory role of miRNAs in health and disease (Calin and Croce, 2006). Deep sequencing (next-generation sequencing) platforms have recently emerged as powerful technologies providing unprecedented insights into biological systems (Metzker, 2010). With automated sequencing technologies becoming routinely used, the data acquisition of millions of sequence reads per experiment leads researchers to the subsequent challenge of data analysis.

*To whom correspondence should be addressed.

Currently, there are several available software packages that facilitate the analysis of small RNA short-read data. Each software is designed via different approaches bearing its advantages and some limitations. Particularly, software packages for miRNA data analysis such as miRDeep (Friedländer *et al*., 2008), do not provide differential expression analysis between known miRNAs in the input samples and do not have a graphic interface. We note that this (and other) software is designated to predict novel miRNAs. Other tools, such as UEA sRNA toolkit (Moxon *et al*., 2008), miRanalyzer, (Hackenberg *et al*., 2009) SeqBuster (Pantano *et al*., 2010), DSAP (Huang *et al*., 2010) and mirTools (Zhu *et al*., 2010), require many processing steps and are mostly web based, thus are either limited in file size or add a long upload stage to the analysis process. This is underscored by the fact that many owners of deep sequencers do not have access to large-scale computing facilities or are not savvy with their use. Finally, none of the existing software provides a method to rigorously deal with multiply mapped reads (Wang *et al*., 2009). Treating multiply mapped reads correctly is essential since small RNAs inherently result in short reads (even shorter than the accustomed 36 nt) and consequently these reads tend to be mapped to multiple genomic locations. Most current software packages for the analysis of small RNA sequencing data remove such reads from the analysis, thus sacrificing the accuracy of any expression profile generated.

Here, we introduce miRNAkey, a software package designed to be used as a base station for the analysis of miRNA sequencing data. miRNAkey can be locally run on any Unix/Linux or Mac computer with 64-bit architecture via the graphical user interface, or on a computer-cluster via the command line. The software is freely available for download at http://ibis.tau.ac.il/miRNAkey.

2 METHODS

MiRNAkey is an intuitive tool, for the implementation of the first steps of analysis of deep sequencing data obtained in miRNA sequencing experiments. The main steps include: (i) searching for and removing the adaptor sequence from the 3′-ends of the reads; (ii) mapping the reads to known miRNA databases, such as miRBase; (iii) counting reads mapped to the different miRNA species in each sample (i.e. sequencing lane), and converting these counts into the normalized RPKM expression-index (reads per kilobase per million mapped reads) to allow comparison across experiments; (iv) quantifying differential expression for miRNAs between paired samples, using chi-squared analysis, thus obtaining *P*-values for the differential expression of miRNAs; (v) generating additional information regarding the input data, such as multiple mapping levels and post-clipping read lengths (see further explanation in Supplementary Data).

An important and unique feature of miRNAkey is the use of the recently developed SEQ-EM algorithm (Pasaniuc *et al*., 2010), for optimizing the distribution of multiply-aligned-reads among the observed miRNAs, rather than discarding them, as is commonly done in this type of analysis. For miRNA data, multiply aligned reads make for a significant fraction of the data, especially for datasets with abundant miRNAs of similar sequence, such as the case in humans, where multiply aligned reads constitute up to 30% of the mapped reads. Accordingly, discarding this portion of the data may lead to a significantly different and biased expression profile.

Differential expression of miRNAs between paired samples is measured using a chi-squared statistic (see Supplementary Data). *P*-values are calculated for the null hypothesis of no differential expression between the two samples. Final *P*-values are corrected using the Bonferroni correction for multiple hypotheses testing.

Deep sequencing was carried out at the Tel Aviv University Genome High-Throughput Sequencing Laboratory following Illumina's Small RNA sample preparation protocol v1.5.

3 RESULTS

The main output generated by miRNAkey is a table comparing each pair of samples/files. This table contains the union of the observed miRNAs in the two samples, ranked from most differentially expressed, to least (see Supplementary Fig. 1). Each line contains the read-counts and RPKM expression indices for a specific miRNA in the two samples, and the differential expression measure for the miRNA between the two samples, along with a *P*-value and additional useful information such as data statistics and multiple read determinations. The ranking of miRNAs allows the user to focus on the most differentially expressed miRNAs in a pair of samples/conditions, thus reducing the high volumes of data generated from deep sequencing experiments. Additionally, many other output reports and plots are generated by miRNAkey (see Supplementary Data), along with files to be used in downstream analysis.

Since miRNAkey is a flexible stand-alone application we found it more appropriate to map the short reads to a known miRNA database (miRbase) rather than to an entire reference genome. This allows a much faster and more flexible analysis with minor compromising in accuracy (see description in Supplementary Data).

We used miRNAkey to evaluate differential expression for a pair of samples. The samples consisted of a viral infected cell line, versus a non-infected control (see Supplementary Data). The exact extracted total RNA was used for deep sequencing analysis and for quantitative real-time PCR. Using miRNAkey, we discovered a number of highly differentially-expressed miRNAs between the two samples, which are likely to be pivotal players in the viral infection process. The most prominent miRNAs were also validated using quantitative real-time PCR and indicated a strong correlation between the methods (see Fig. 1).

4 CONCLUSION

miRNAkey is a software package designed as a base station for the analysis of miRNA sequencing data. A friendly graphical user interface is applied to determine the analysis steps. The tabular output contains general and detailed reports on the sequence reads and provides an accurate picture of the differentially expressed miRNAs in paired samples. miRNAkey can run on a local Unix/Linux or Mac computer and deals with multiply mapped miRNA reads. The software can be downloaded at http://ibis.tau.ac.il/miRNAkey.

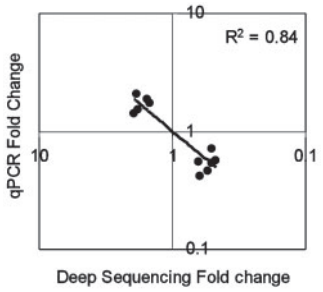

Fig. 1. A comparison of miRNAkey identified differentially regulated miRNAs versus qPCR validated miRNA expression changes.

ACKNOWLEDGEMENTS

We thank Alon Shtivelman, Dr Noah Zaitlen and Dr Bogdan Pasaniuc for technical assistance. We thank Dr Varda Oron-Karni and Dr Orly Yaron from the Tel Aviv University Genome High-Throughput Sequencing Laboratory for their dedicated and professional work on the Illumina Genome Analyzer IIx.

Funding: Chief Scientist Office, Ministry of Health, Israel; The Kurz-Lion Foundation; The Tel Aviv University, Faculty of Medicine, Schreiber Foundation; The Wolfson Family Charitable Fund (to N.S.); Israeli Science Foundation (04514831); the National Science Foundation (IIS-071325412 to E.H.); Faculty fellow of the Edmond Safra Bioinformatics program at Tel-Aviv University (to E.H.); Edmond Safra Bioinformatics Program at Tel Aviv University and the Israeli Science Foundation (04514831) to R.R. in part.

Conflict of Interest: none declared.

REFERENCES

Bushati,N. and Cohen,S.M. (2007) microRNA functions. *Annu. Rev. Cell Dev. Biol.*, **23**, 175.

Calin,G.A. and Croce,C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857–866.

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

Golan,D. *et al*. (2010) Biased hosting of intronic microRNA genes. *Bioinformatics*, **26**, 992–995.

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

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

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

Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet*., **11**, 31–46.

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

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.

Pasaniuc,B. *et al*. (2010) Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *Proceedings of Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010*, Lisbon, Portugal, August 12–15, pp. 397–407.

Shomron,N. *et al*. (2009) An evolutionary perspective of animal microRNAs and their targets. *J. Biomed. Biotechnol*., **2009**, 594738.

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

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