BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 15 2012, pages 2059–2061
doi:10.1093/bioinformatics/bts311

Sequence analysis
Advance Access publication May 24, 2012

The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets

Matthew B. Stocks[1,†], Simon Moxon[2,†], Daniel Mapleson[1], Hugh C. Woolfenden[1], Irina Mohorianu[1], Leighton Folkes[1], Frank Schwach[3], Tamas Dalmay[4] and Vincent Moulton[1,*]

[1]School of Computing Sciences, University of East Anglia, Norwich Research Park, NR4 7TJ, Norwich, UK
[2]Department of Genetics, Yale University, School of Medicine, 333 Cedar Street, New Haven CT 06520, USA
[3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK and
[4]School of Biological Sciences, University of East Anglia, Norwich Research Park, NR4 7TJ, Norwich, UK

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: RNA silencing is a complex, highly conserved mechanism mediated by small RNAs (sRNAs), such as microRNAs (miRNAs), that is known to be involved in a diverse set of biological functions including development, pathogen control, genome maintenance and response to environmental change. Advances in next generation sequencing technologies are producing increasingly large numbers of sRNA reads per sample at a fraction of the cost of previous methods. However, many bioinformatics tools do not scale accordingly, are cumbersome, or require extensive support from bioinformatics experts. Therefore, researchers need user-friendly, robust tools, capable of not only processing large sRNA datasets in a reasonable time frame but also presenting the results in an intuitive fashion and visualizing sRNA genomic features. Herein, we present the UEA sRNA workbench, a suite of tools that is a successor to the web-based UEA sRNA Toolkit, but in downloadable format and with several enhanced and additional features.

Availability: The program and help pages are available at http://srna-workbench.cmp.uea.ac.uk.

Contact: vincent.moulton@cmp.uea.ac.uk

Received on February 10, 2012; revised on May 8, 2012; accepted on May 22, 2012

1 INTRODUCTION

RNA silencing is a complex, highly conserved mechanism mediated by small RNAs (sRNAs) (Chapman and Carrington, 2007). sRNAs, such as micro RNAs (miRNAs), are typically 20–24 nt in length and act as guide molecules to regulate gene expression. Next generation sequencing (NGS) technologies have become the *de facto* way to generate sRNA datasets. These datasets often come from several samples, each consisting of millions of distinct reads. Therefore, there is an increasing need for computational tools for exploring and analysing such data.

Several tools have been developed for NGS sRNA analysis (see Li *et al.* (2012) for a recent overview of tools for miRNA analysis). These are typically command-line driven, focused on a specific application (e.g. miRNA detection) and often require installation of supporting software packages. In addition, many of them lack tools for visualizing results.

Herein, we present the UEA sRNA workbench, a suite of interactive Java-based sRNA analysis tools, which provides an easy-to-use, well-documented platform to create workflows for processing sRNA NGS data. The workbench is a successor to the web-based UEA sRNA Toolkit (Moxon *et al.*, 2008), which, since its establishment in 2009, has received an average of 12 000 page views per month and processed over 5400 sRNA datasets. As well as being downloadable, the new workbench provides many improved features compared with the toolkit, with new tools for visualizing data and processing multiple datasets. In addition, the workbench is limited only by local memory whereas the toolkit had strict datasize limits due to the nature of its web-based interfaces. The workbench can also be command-line driven to allow users to easily plug it into existing pipelines (Studholme, 2011).

2 DESCRIPTION OF THE WORKBENCH

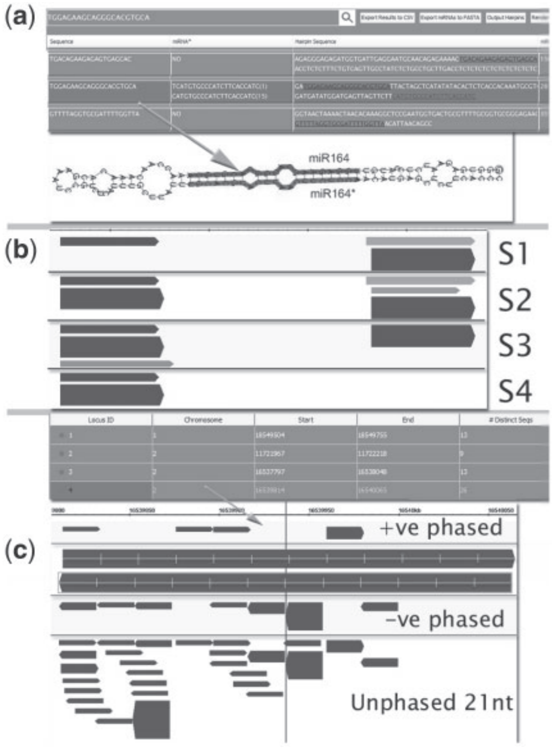
2.1 Helper tools

- *Adaptor remover* removes adaptor fragments from raw short read sequence data and outputs data to FASTA format.
- *Filter* produces a filtered version of an sRNA dataset, controlled by several user-defined criteria, including sequence length, abundance, complexity, transfer and ribosomal RNA removal.
- *Sequence alignment* provides a graphical front-end for the PatMaN (Prüfer *et al.*, 2008) sequence alignment tool.

2.2 Computational analysis tools

- *miRCat (miRNA categorization)* predicts mature miRNAs and their precursors from an sRNA dataset and a genome.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
[*]To whom correspondence should be addressed.



TGGAGAAGCAGGGCACGTGCA

Sequence	miRNA*	Hairpin Sequence
TGACAGAAGAGAGTGAGCAC	NO	AGAGGCAGAGATGGTGATTGAGGAATGCAACAGAGAAAAC[illegible] ACCTCTCTTTCTGTCAGTTGCCTATCTCTGCCTGCTTGACCTCTCTCTCTCTCTCTCTC

TGGAGAAGCAGGGCACGTGCA	TCATGTGCCCATCTTCACCATC(1)
	CATGTGCCCATCTTCACCATC(15)
		GA[illegible]TTACTAGCTCATATATACACTCTCACCACAAATGCGT
		GATGATATGGATGAGTTACTTCTT[illegible]

GTTTTAGGTGCGATTTTGGTTA	NO	GGTAACTAAAACTAACACAAAGGCTCCCAATGGTGACTGCGTTTTGCGGTGCCGGAGAAC [illegible]ACATTAACAGCC

Locus ID	Chromosome	Start	End	# Distinct Seqs
1	1	18549504	18549755	13
2	2	11721967	11722218	9
3	2	16537797	16538048	13
4	2	16539814	16540065	26

Fig. 1. Workflow examples from the UEA sRNA workbench (data from Rajagopalan *et al.* (2006)). (**a**) After miRCat has classified miR164, a secondary structure plot is generated through the Hairpin Annotation tool. (**b**) After SiLoCo predicts the miR164 locus in four sRNA samples (including the sample used in (a)), output is sent to VisSR for visualization. Reads of lengths 19, 20–21 and 22–23 are coloured pink, red and green, respectively. (**c**) ta-siRNA prediction of a TAS gene. The locus is visualized using VisSR

The algorithm uses the technique in Moxon *et al.* (2008) for predicting miRNA precursor hairpins. The algorithm has been enhanced to handle animal datasets and to use multi-threading hardware, improving run-time performance. Users are presented with a table containing details of predicted miRNA candidates and precursors. miRNA candidates are reported as they are discovered allowing users to immediately begin down-stream analysis. miRNAs in miRBase (Griffiths-Jones, 2010) are also reported as such.

- *SiLoCo (short interfering RNA locus comparison)* compares sRNA expression levels by grouping sRNAs into loci based on genomic location (Moxon *et al.*, 2008). SiLoCo has been extended to allow comparison of multiple samples. The output is displayed in a table containing normalized abundance of the sequences and a weighted count on the number of occurrences in the genome. In addition, the mean abundance score for the locus and the maximum fold change between normalized abundances over each sample are reported.
- *ta-siRNA (trans-acting short interfering RNA) prediction* is an implementation of an algorithm proposed in Chen *et al.* (2007) for prediction of phased ta-siRNAs in plant sRNA datasets. Users are presented with the coordinates of the TAS loci and the phased sRNAs (Fig. 1c).
- *miRProf (miRNA profiler)* determines normalized expression levels of sRNAs matching known miRNAs in miRBase. This tool has been enhanced for comparison of miRNA expression levels across multiple samples. Users are presented with a table containing raw, normalized and weighted counts of miRNAs. Users can also combine results through several criteria including miRNAs containing mismatches, miRNAs found in different organisms and miRNA variants.

2.3 Visualization tools

- *Hairpin annotation* generates a secondary structure from an RNA sequence and highlights regions of interest using RNAplot (Hofacker *et al.*, 1994). This tool can be activated from miRCat to display predicted miRNA precursors with both the miRNA and the miRNA* highlighted (Fig. 1a). Alternatively, a user can input long and up to 14 short sequences to generate a secondary structure. For each displayed miRNA precursor, the minimum free energy of the structure is also reported.
- *VisSR (visualization of sRNAs)* Visualization of sRNA features can be used to gain insight into their likely biogenesis or function. To do this, VisSR uses components from GenoViz (Helt *et al.*, 2009) to generate a visual representation of sRNAs and user-imported genomic features. Arrows are used to depict the strand to which a short read maps to, colour is used to represent the size class of the short read and thickness illustrates abundance. VisSR can be activated from various tools to display predicted sRNA loci with each sample displayed in separate tiers (Fig. 1b), phased sRNAs for both strands in separate tiers and all unphased 21nt sRNAs in a further tier (Fig. 1c), and predicted miRNAs with miRNA* if present with the precursor as it appears on the genome. Output is generated in an image format, suitable for publication.

3 DISCUSSION

The relative low cost of NGS has generated an abundance of sRNA data necessitating a demand for dedicated bioinformatics support. The sRNA workbench enables biologists to gain an understanding of the sRNAome in plants and animals through the efficient processing of large NGS sRNA datasets, complemented with easily accessible visualization tools. It is hoped that these tools will enable biologists to shed light on the key molecular pathways involved in RNA silencing.

Funding: Biotechnology and Biological Sciences Research Council (grant BBSRC BB/I00016X/1).

Conflict of Interest: none declared.

REFERENCES

Chapman, E.J., and Carrington, J.C. (2007) Specialization and evolution of endogenous small RNA pathways. *Nat. Rev. Genet*., **8**, 884–896.

Chen,H.-M. *et al.* (2007) Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proc. Natl. Acad. Sci. USA*, **104**, 3318–3323.

Griffiths-Jones,S. (2010) miRBase: microRNA sequences and annotation. *Curr. Protoc. Bioinformatics*, **Chapter 12**, Unit 12.9.1–Unit 12.910.

Helt,G.A. *et al.* (2009) Genoviz software development kit: Java tool kit for building genomics visualization applications. *BMC Bioinformatics*, **10**, 266.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemi. Mon.*, **125**, 167–188.

Li,Y. *et al.* (2012) Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Res*., **40**, 4298–4305.

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

Prüfer,K. *et al.* (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, **24**, 1530–1531.

Rajagopalan,R. *et al.* (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev*., **20**, 3407–3425.

Studholme,D.J. (2012) Deep sequencing of small RNAs in plants: applied bioinformatics. *Brief Funct. Genom*., **11**, 71–85