BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 19 2008, pages 2252–2253
doi:10.1093/bioinformatics/btn428

*Genome analysis*

# A toolkit for analysing large-scale plant small RNA datasets

Simon Moxon[1,†], Frank Schwach[1,†], Tamas Dalmay[2], Dan MacLean[3], David J. Studholme[3] and Vincent Moulton[1,*]

[1]School of Computing Sciences, [2]School of Biological Sciences, University of East Anglia, Norwich, NR4 7TJ and [3]The Sainsbury Laboratory, Colney Lane, Norwich, NR4 7UH, UK

Received on May 27, 2008; revised on July 14, 2008; accepted on August 10, 2008

Advance Access publication August 19, 2008

Associate Editor: Ivo Hofacker

**ABSTRACT**

**Summary:** Recent developments in high-throughput sequencing technologies have generated considerable demand for tools to analyse large datasets of small RNA sequences. Here, we describe a suite of web-based tools for processing plant small RNA datasets. Our tools can be used to identify micro RNAs and their targets, compare expression levels in sRNA loci, and find putative trans-acting siRNA loci.

**Availability:** The tools are freely available for use at http://srna-tools.cmp.uea.ac.uk

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

## 1 INTRODUCTION

Several classes of small (20–30 nt) non-coding RNAs (sRNAs) can be distinguished by biogenesis and function in post-transcriptional gene regulation and epigenetic control in plants, animals and fungi (for reviews see: Brodersen and Voinnet, 2006; Lippman and Martienssen, 2004). Micro RNAs (miRNAs) and trans-acting siRNAs (ta-siRNAs) are two important classes of sRNAs that both induce post-transcriptional silencing of target genes. Computationally, miRNAs can be identified by their characteristic fold-back precursors, while ta-siRNA are found by a 'phased' alignment pattern at their genomic regions of origin (Axtell *et al*., 2006).

Novel high-throughput sequencing technologies greatly facilitate small RNA detection and analysis (Hafner *et al*., 2007). However, the lack of supporting data analysis tools presents a major bottleneck. Here, we present an easy-to-use web-based toolkit that is specifically geared towards the analysis of large-scale plant sRNA datasets. Plant specific tools are necessary due to important differences in the biogenesis and mode of action between plant and animal sRNAs (Millar and Waterhouse, 2005).

## 2 DESCRIPTION OF THE TOOLS

### 2.1 miRCat: miRNA detection

miRCat identifies mature miRNAs and their precursors. Users upload a FASTA file of sRNA sequences, which are mapped to a plant genome using PatMaN (Prüfer *et al*., 2008) and grouped into loci. To enrich for miRNA candidates, a number of empirical and published criteria for bona fide miRNA loci are applied by the software (Jones-Rhoades *et al*., 2006, details listed on the tool's website). In brief, the program searches for a two-peak alignment pattern of sRNAs on one strand of the locus and assesses the secondary structures of a series of putative precursor transcripts using the RNAfold (Hofacker *et al*., 1994) and randfold (Bonnet *et al*., 2004) programs. As a result, miRCat produces three files: (i) a comma-separated text (csv) file with the details for predicted miRNA candidates, (ii) the RNAfold output for candidate precursors and (iii) a FASTA file of predicted mature miRNA sequences. miRCat has been tested on several high-throughput plant sRNA datasets and shows a high level of sensitivity and specificity. When tested on a publicly available *Arabidopsis* leaf sRNA dataset (GEO accession GSM118373; Rajagopalan *et al*., 2006) containing 186 899 sRNA sequences, miRCat predicted 89 miRNA loci using default parameters. Eighty-three of these predictions were known miRNA sequences and 6 novel miRNA loci were predicted (Fig. 1a). There were 91 known miRNA loci with an sRNA abundance of five or more (default threshold for miRCat) in the dataset. This shows 91.2% sensitivity and, even if all novel predictions would have been false positives, this would give a specificity of 99.93% (8362 loci tested). As a web-based tool, miRCat complements related software developed for local installation and command line use, such as a recently published program for discovering miRNAs in animal datasets (Friedländer *et al*., 2008).

### 2.2 SiLoCo: sRNA locus expression comparison

High-throughput sequencing can be used to compare sRNA expression profiles under varying conditions or between mutants and wild-type to gain insights into the biogenesis and function of sRNAs. Plant sRNA populations are highly complex with many genomic loci producing highly diverse sRNA populations. In such cases, individual sequences may not be found more than once even in very large datasets, thus making it necessary to group sRNAs by their locus of origin in the genome and compare expression levels on a locus, rather than individual sequence levels. Such an approach also needs to take into account the degree of repetitiveness of sRNA matches to the genome. SiLoCo identifies sRNA loci on plant genomes from two sRNA datasets, which can be uploaded by the user and/or selected from publicly available datasets. SiLoCo maps sRNA sequences to the genome using PatMaN (Prüfer *et al*., 2008)

*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.

 

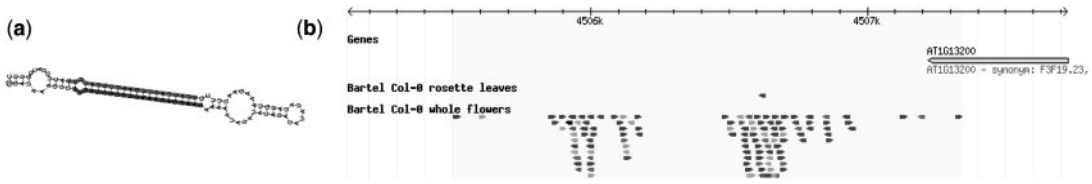


**Fig. 1.** Example results: (**a**) secondary structure plot of a putative novel miRNA identified by miRCat in a publicly available dataset (Rajagopalan *et al*., 2006). MiRNA/miRNA* highlighted in red/purple using the RNAfold/annotate tool on our website. (**b**) The top-ranking differentially expressed locus found by SiLoCo in a comparison of two public Arabidopsis sRNA datasets (leaf and flower tissue; Rajagopalan *et al*., 2006) shown in the ASRP genome browser (Backman *et al.*, 2008). Identified sRNA locus highlighted in yellow, sRNA matches shown by coloured arrows. Genome browser tracks: genes (top), leaf sRNAs (middle), flower sRNAs (bottom).

and weighs each sRNA hit by its repetitiveness in the genome. Loci are defined as described previously (Molnár *et al*., 2007; Mosher *et al*., 2008) by a minimum number of sRNA hits to a region and a maximum 'gap', i.e. absence of sRNA hits, between them. Hit counts are normalized to the total number of genome-matching reads in each sample to make them comparable. For each locus, the log2 ratio and the average of the normalized sRNA hit counts are calculated and ranked independently. A sum of the two ranks is also provided and the results can be downloaded as a csv-formatted file. Sorting the list of loci by the rank sum in a spreadsheet program is an easy way of finding the best candidates for differentially expressed loci, where sRNA abundance differs greatly at a high overall expression level (Fig. 1b). Hyperlinks to some public genome browsers can also be included in the result file.

### 2.3 ta-siRNA prediction

ta-siRNAs are produced from a double-stranded RNA molecule. Alignments of ta-siRNAs to their region of origin exhibit a characteristic 'phased' pattern (Axtell *et al*., 2006) that can be identified computationally. Our tool is a web-based implementation of an algorithm proposed by Chen *et al*. (2007) for calculating the probability of obtaining the observed percentage (or more) of phased sRNA matches by chance. An adjustable *P*-value cutoff is used to filter for loci with a significant degree of 21 nt phasing. Results are downloadable as a csv file. A test run with a publicly available Arabidopsis dataset (Rajagopalan *et al*., 2006) returned eight candidate loci, including four known ta-siRNA loci and three phased loci also reported by Chen *et al*. (2007).

### 2.4 Helper tools

We provide a web tool to find target transcripts of sRNAs based on published rules for plant miRNAs (Allen *et al*., 2005; Schwab *et al*., 2005). This tool allows batch searching of up to 50 sRNAs against 20 different plant gene datasets. In addition, we provide an interface to the RNAfold/RNAplot programs (Hofacker *et al*., 1994) that allows the visualization of miRNA candidates. This tool accepts a precursor RNA and sRNA sequences which are highlighted on the resulting secondary structure (Fig. 1a).

## 3 DISCUSSION

High-throughput sRNA sequencing has great potential to identify new members of known sRNA classes, especially in tissues or under environmental conditions that have not been investigated yet. The technology can also be used to compare sRNA profiles, thus gaining further insights into sRNA biogenesis and function. Our tools are ideally suited for these types of analyses on plant sRNA data and are easy to use.

## ACKNOWLEDGEMENTS

The authors wish to thank D.C. Baulcombe and K. Kelly for helpful ideas and discussions and A. Courtenay, C. Collins and M. Burrell for IT support.

*Funding*: This work was supported by the Biotechnology and Biological Sciences Research Council [grant number BB/E004091/1] and the Gatsby Charitable Foundation (to D.M. and D.J.S.).

*Conflict of Interest*: none declared.

## REFERENCES

Allen,E. *et al*. (2005) microRNA-directed phasing during transacting siRNA biogenesis in plants. *Cell*, **121**, 207–221.

Axtell,M.J. *et al*. (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell*, **127**, 565–577.

Backman,T.W. *et al.* (2008) Update of ASRP: the Arabidopsis small RNA project database. *Nucleic Acids Res.*, **36**, D982–D985.

Brodersen,P. and Voinnet,O. (2006) The diversity of RNA silencing pathways in plants. *Trends Genet*., **22**, 268–280.

Bonnet,E. *et al*. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.

Chen,H. *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.

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

Hafner,M. *et al.* (2007) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, **44**, 3–12.

Hofacker,I.L. *et al*. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Jones-Rhoades,M.W. *et al*. (2006) MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, **57**, 19–53.

Lippman,Z. and Martienssen,R. (2004) The role of RNA interference in heterochromatic silencing. *Nature*, **431**, 364–370.

Millar,A. and Waterhouse,P.M. (2005) Plant and animal microRNAs: similarities and differences. *Funct. Integr. Genomics*, **5**, 129–135.

Molnár,A. *et al*. (2007) miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, **447**, 1126–1129.

Mosher,R.A. *et al*. (2008) PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc. Natl Acad. Sci. USA*, **105**, 3145–3150.

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.

Schwab,R. *et al*. (2005) Specific effects of microRNAs on the plant transcriptome. *Dev. Cell*, **8**, 517–527.