Published online 14 December 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D829–D833***
doi:10.1093/nar/gkl991

CSRDB: a small RNA integrated database and browser resource for cereals

Cameron Johnson[1,*], Lewis Bowman[2], Alex T. Adai[3], Vicki Vance[2] and Venkatesan Sundaresan[1,4]

[1]Section of Plant Biology, College of Biological Sciences, University of California, Davis, CA 95616, USA, [2]Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA, [3]Biological and Medical Informatics, University of California, San Francisco, CA 94143, USA and [4]Department of Plant Sciences, College of Agricultural Sciences, University of California, Davis, CA 95616, USA

Received August 15, 2006; Revised October 26, 2006; Accepted October 30, 2006

ABSTRACT

Plant small RNAs (smRNAs), which include microRNAs (miRNAs), short interfering RNAs (siRNAs) and *trans*-acting siRNAs (ta-siRNAs), are emerging as significant components of epigenetic processes and of gene networks involved in development and in homeostasis. Here we present a bioinformatics resource for cereal crops, the Cereal Small RNA Database (CSRDB), consisting of large-scale datasets of maize and rice smRNA sequences generated by high-throughput pyrosequencing. The smRNA sequences have been mapped to the rice genome and to the available maize genome sequence and these results are presented in two genome browser datasets using the Generic Genome Browser. Potential RNA targets for the smRNAs have been predicted and access to the resulting smRNA/RNA target pair dataset has been made available through a MySQL based relational database. Various ways to access the data are provided including links from the genome browser to the target database. Data linking and integration are the main focus for this interface, and internal as well as external links are present. The resource is available at http://sundarlab.ucdavis.edu/smrnas/ and will be updated as more sequences become available.

INTRODUCTION

MicroRNAs (miRNAs), small interfering RNAs (siRNAs) and *trans*-acting siRNAs (ta-siRNAs) are small RNAs (smRNAs) of ~19–24 nt that act as important negative regulators of genes and other nucleotide sequences [for recent reviews, see (1–4)]. Both miRNAs and ta-siRNAs have been implicated in the regulation of genes involved in development and homeostasis. SiRNAs are important suppressors of transposons and viruses, but are also implicated in processes of homeostasis as well as in the maintenance of epigenetic states such as those in heterochromatic and centromeric regions of the genome.

The classification of a smRNA as a miRNA, siRNA or ta-siRNA depends largely on the biogenesis and mode of action of the smRNA. MiRNAs are processed from an incompletely base-paired region of a folded RNA molecule and act in *trans* on RNA transcripts synthesized from other regions of the genome. SiRNAs primarily act on the same RNA molecule from which they derive: they are processed from fully double-stranded RNA that arises via transcription of hairpin transgenes, from the activity of an RNA dependent RNA polymerase on an RNA template or from *cis*- and *trans*-natural antisense transcripts. Ta-siRNAs are sets of phased smRNAs that derive from a fully double stranded RNA that arises by the activity of an RdRP. Like miRNAs, the ta-siRNAs act in *trans* to negatively regulate target transcripts from other loci.

Here we present an interface to a preliminary smRNA dataset along with potential mRNA targets. The smRNA sequences obtained using high-throughput pyrosequencing (5) by 454 Life Science have been mapped to the complete rice genome and a partial maize genome and are presented within two genome browsers enabling the potential sources of these smRNAs and their local genomic relationships to be identified. The genome browsers represent one interface to a relational database of potential mRNA targets predicted using the FASTH program, and other ways to search the data are also provided.

THE SMALL RNA SEQUENCES AND THE GENOME BROWSERS

The rice smRNA library was constructed from a mixture of RNA isolated from 30 to 60 day leaves (~16.5% each), 10, 25 and 30 day seedlings (~11% each), 4–7 cm inflorescences

*To whom correspondence should be addressed. Tel: +1 530 754 9852; Fax: +1 530 752 5410; Email: camjohnson@ucdavis.edu

(~16.5%) and 25 day seedling polysomes (~16.5%). The maize smRNA library was constructed from a mixture of RNA from 7 day seedlings (~10%), adult, juvenile and embryonic leaves (~10% each), immature ears, 2–5 cm (25%), immature tassels, 3–5 cm (25%) and stems (10%). SmRNAs were ligated to adapters and amplified as described (6). The amplified molecules were sequenced using high-throughput pyrosequencing by 454 Life Science, a procedure for short sequence reads.

A total of 92 298 rice and 227 710 maize smRNA sequences were obtained (Table 1). The sequences of the primers that were used for amplifying the ligated molecules were used to identify the termini of the smRNA sequences that were in turn selected for sizes from 18 to 34 nt resulting in 54 111 and 158 581 accepted sequences from rice and maize, respectively. The ligation method used enables the polarity information for the smRNA sequences to be retained. These sequences were then mapped without allowing for any mismatches, using a hash table lookup method implemented in a perl script, to the OSA1 TIGR release four rice genome sequence (7) and the available MAGI sequence contigs (8–10). Of the 54 111 rice sequences, 35 454 could be mapped to the genome, while of the 158 581 maize sequences, 68 871 could be mapped. These mapped sequences correspond to 12 819 and 26 070 unique sequences, respectively, and are available for download at http://sundarlab.ucdavis.edu/smrnas/data/.

Table 1. Small RNA sequence count

Species	Total 454 reads	Extracted and accepted	Mapped zero mismatches	
			Total	Unique
Maize	227 710	158 581	68 871	26 070
Rice	92 298	54 111	35 454	12 819

The mapped smRNAs for both rice and maize are provided within an implementation of the generic genome browser (11) and have been grouped into tracks corresponding to individual size classes 18–24 nt (Figure 1). Within plants smRNAs >24 nt are not yet known to be biologically significant, but these 25–34 nt smRNAs are included in the browser as an additional pooled track. From each smRNA annotation, a link leads to a predicted set of mature gene transcript targets for rice and maize (see below for a detailed description). Sequences for the miRNAs and miRNA-precursors (miRBase—release 8.2) maintained at Rfam by the Sanger Institute (12,13) have also been mapped to the respective genome sequences. In addition, TIGR gene ontology annotations for the rice genes are provided from the TIGR website using

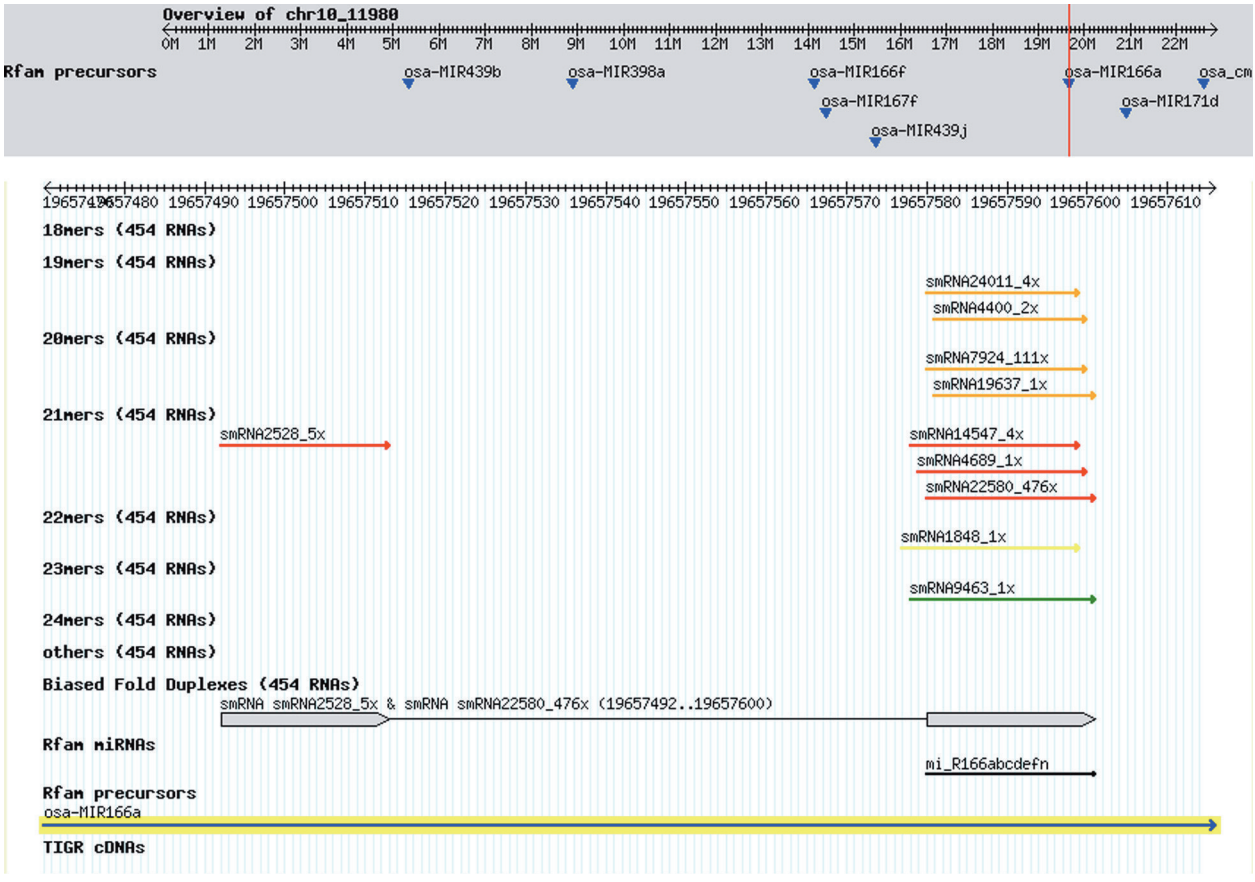

Figure 1. Genome browser view of smRNA-gbrowse for the rice MIR166a locus. The Rfam small RNA sequence for miR166a is the same for the seven miRNA loci MIR166 a, b, c, d, e, f and n. The most frequent 454 sequences for miR166a are 20 and 21 nt. These occur 111 and 476 times, respectively, as indicated within the name of the smRNAs (i.e. smRNA7924_111x, smRNA22580_476x). A bioinformatically predicted duplex of two smRNAs, smRNA2528_5x and smRNA22580_476x that represent the miRNA* and miRNA sequences, respectively, is shown in the track called 'Biased Fold Duplexes'.

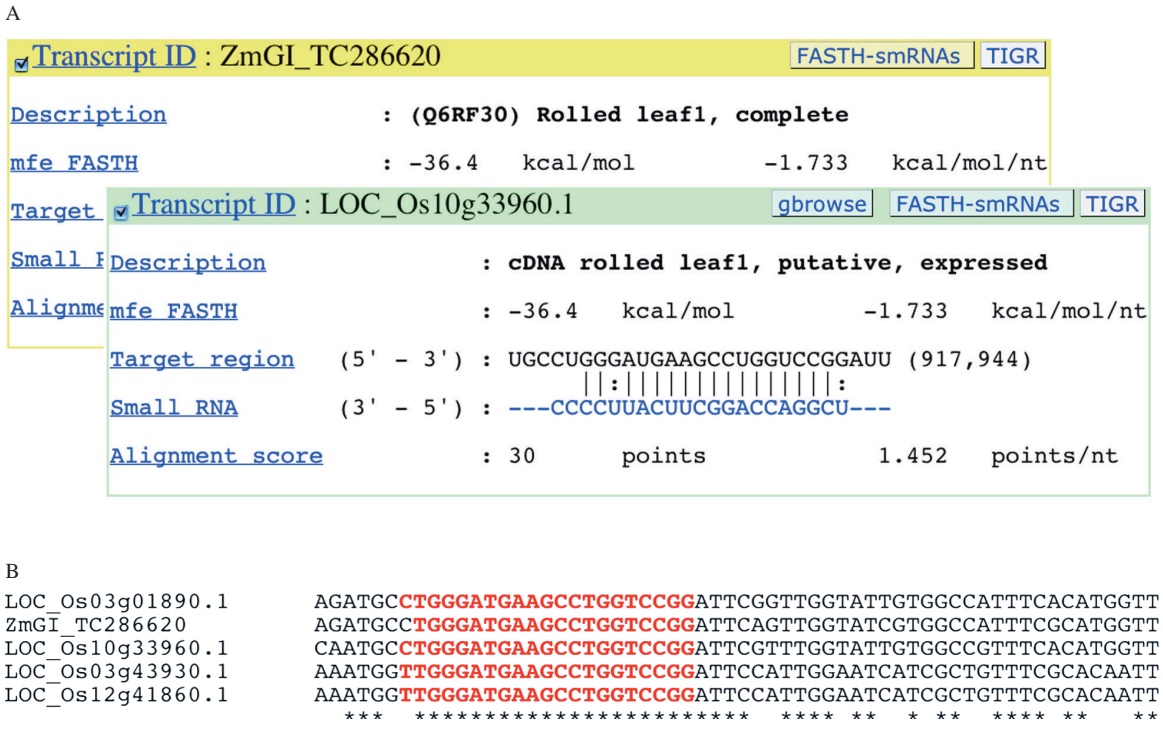

Figure 2. Accessing the target transcript database via a smRNA ID query or from a smRNA within the genome browser returns a results page of potential target transcripts for both rice and maize. (**A**) Predicted target records contain the transcript ID, a brief description from the TIGR annotation, an estimated thermal stability for the smRNA-target duplex, the aligned target region with the small RNA sequence, and an alignment score. SmRNA length normalized scores are shown to the right in the thermal stability and alignment score lines. These records are ranked according to the normalized alignment score and then by the normalized thermal stability. (**B**) Related transcripts may be selected using the check boxes. When submitted, an alignment of the checked sequences and the annotated target sites will be generated to enable easy detection of conserved target sites.

the distributed annotation system (DAS) (14). The rice browser also contains tracks for pairs of adjacent smRNAs that are capable of base pairing to form smRNA duplexes consistent with processing from single RNA molecules with secondary structures as in the case of miRNAs.

POTENTIAL TARGETS DATABASE

The rice and maize smRNAs that could be mapped to their respective genome sequences were used to search for potential smRNA target sites within the 62 827 rice and 36 563 maize mature mRNA transcript sequences from TIGR. This was performed using the FASTH software of Zuker (15) that returns results based on estimated thermal stabilities. In a post-processing step these predicted smRNA target pairs were reformatted and alignment scores provided to enable appropriate ranking of the predicted smRNA-target duplexes. The alignment scores were generated using a position dependent Smith-Waterman based scheme with greater penalties for mismatches, GU pairs and bulges within the 2–13 nt inclusively in a similar way as described previously (16,17) and reduced penalties in the end nucleotides of the smRNA (see the website for the current scoring parameters). These data are provided in a MySQL based relational database that is linked to both the rice and maize genome browser interfaces. The smRNA-target duplex database can also be queried independently of the genome browser using mature mRNA transcript IDs for rice and maize or the assigned smRNA IDs.

RESULT FORMATS AND INTEGRATION

When a link from a smRNA within the genome browser is followed, a page of potential target transcripts is returned that are ranked based on the alignment score normalized for smRNA length and then by thermal stability also normalized for smRNA length. Included in each predicted target record, is a description of the potential target gene (where available), the thermal stability of the smRNA-target duplex as estimated by FASTH, the aligned target and smRNA sequences, the alignment score and the length normalized values (Figure 2A). Conserved target sites may be identified within different genes that belong to the same family by selecting the sequences using the provided checkboxes. When submitted, these sequences will be aligned and the single best target site from each transcript will be annotated in boldface red type (Figure 2B), enabling conserved target sites to be readily visualized. In addition, following a link on the predicted targets page for any target RNA returns a list and a graphic of smRNAs that may target this single transcript (Figure 3). The length normalized alignment scores are indicated by the intensity of the smRNA boxes in the graphic. The mRNA target database can also be searched from the main website interface. SmRNA queries and transcript ID queries will return results pages as described, respectively.

OTHER INTEGRATED INFORMATION

The results pages also contain other relevant links. From a predicted target record, a link exists to TIGR gene ontology

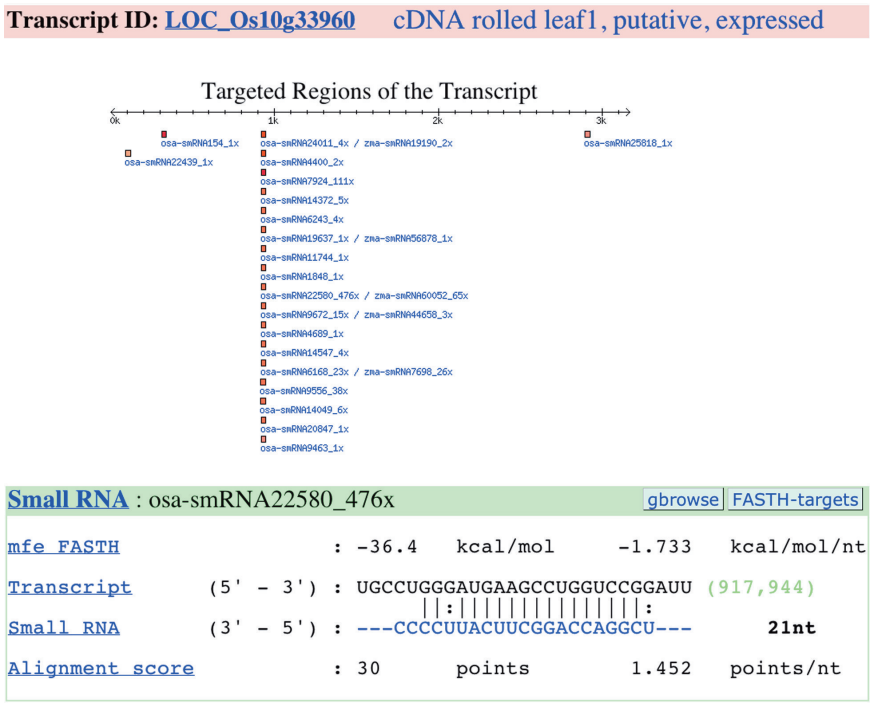

Figure 3. Accessing the target transcript database via a Transcript ID or from within the results page of a smRNA target transcript database query, returns a results page of smRNAs that may target the query transcript. At the top of the page is a graphic showing the distribution of these potential targeting-smRNAs with intensities relating length normalized alignment score. Below this are individual records for each smRNA showing the estimated thermal stability of the duplex, the alignment of the smRNA sequence with the targeted region of the target transcript, and the alignment score. These records are ordered according to their position on the graphic.

information specific for this gene ('TIGR' button). For results showing smRNAs targeting a single transcript, links are provided for each smRNA back to the smRNA browser ('gbrowse' button) and another link that leads back to a separate predicted targets page ('FASTH-targets' button). The main website interface also contains a link to a miRNA knowledge page for rice and maize. This contains current information regarding miRNAs and reports of their target prediction and validation within the species. In addition, each miRNA has links to the smRNA genome browser interface enabling easy browsing of known miRNAs within the genome context.

SMALL RNA TOOLS

On the main website interface page are a number of tools to facilitate analysis of smRNAs, including a BLAST service for performing searches against the sets of smRNA sequences, a tool for looking for conserved target sites within a set of related mature transcripts, and a tool for identifying potential target sites within a single mature transcript.

FUTURE ADVANCES TO THE DATABASE

The smRNA data provided in this release represents a preliminary dataset. Future datasets will include smRNAs isolated from different tissues and under different conditions. These datasets will be useful for the identification of smRNAs that may be expressed under specialized conditions. The tissue-specific differential expression may enable identification of biologically important smRNAs that may not otherwise be distinguished from background siRNAs. In addition to updates in the data, continued improvements in the integration of the website interface are anticipated.

ACKNOWLEDGEMENTS

We would like to thank Micheal Zuker for the use of the FASTH program and Virginia Walbot for providing expertise and assistance in the collection of maize tissues for small RNA isolation. This work was supported by NSF Plant Genome grant #0501760 to V.V., L.B. and V.S. Funding to pay the Open Access publication charges for this article was provided by NSF Plant Genome grant #0501760 to V.V., L.B. and V.S.

Conflict of interest statement. None declared.

REFERENCES

1. Meins,F., Jr, Si-Ammour,A. and Blevins,T. (2005) RNA silencing systems and their relevance to plant development. *Annu. Rev. Cell Dev. Biol.*, **21**, 297–318.
2. Willmann,M.R. and Poethig,R.S. (2005) Time to grow up: the temporal role of small RNAs in plants. *Curr. Opin. Plant Biol.*, **8**, 548–552.
3. Mallory,A.C. and Vaucheret,H. (2006) Functions of microRNAs and related small RNAs in plants. *Nature Genet.*, **38**, S31–S36.
4. Jones-Rhoades,M.W., Bartel,D.P. and Bartel,B. (2006) MicroRNAS and Their Regulatory Roles in Plants. *Annu. Rev. Plant Biol.*, **57**, 19–53.

5. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
6. Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
7. Yuan,Q., Ouyang,S., Wang,A., Zhu,W., Maiti,R., Lin,H., Hamilton,J., Haas,B., Sultana,R., Cheung,F. *et al*. (2005) The Institute for Genomic Research Osa1 Rice Genome Annotation Database. *Plant Physiol*., **138**, 18–26.
8. Whitelaw,C.A., Barbazuk,W.B., Pertea,G., Chan,A.P., Cheung,F., Lee,Y., Zheng,L., van Heeringen,S., Karamycheva,S., Bennetzen,J.L. *et al*. (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science*, **302**, 2118–2120.
9. Palmer,L.E., Rabinowicz,P.D., O'Shaughnessy,A.L., Balija,V.S., Nascimento,L.U., Dike,S., de la Bastide,M., Martienssen,R.A. and McCombie,W.R. (2003) Maize genome sequencing by methylation filtration. *Science*, **302**, 2115–2117.
10. Fu,Y., Emrich,S.J., Guo,L., Wen,T-J., Ashlock,D.A., Aluru,S. and Schnable,P.S. (2005) Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. *Proc. Natl Acad. Sci. USA*, **102**, 12282–12287.
11. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
12. Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res*., **32**, D109–D111.
13. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., **34**, D140–D144.
14. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
15. Zuker,M. (2003) Predicting nucleic acid hybridization and melting profiles. *Genome Inform*., **14**, 266–268.
16. Allen,E., Xie,Z., Gustafson,A.M. and Carrington,J.C. (2005) MicroRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, **121**, 207–221.
17. Schwab,R., Palatnik,J., Riester,M., Schommer,C., Schmid,M. and Weigel,D. (2005) Specific effects of microRNAs on the plant transcriptome. *Dev. Cell*, **8**, 517–527.