the plant journal

SEB Society for Experimental Biology

The Plant Journal (2012) doi: 10.1111/j.1365-313X.2012.04922.x

TECHNICAL ADVANCE

SoMART: a web server for plant miRNA, tasiRNA and target gene analysis

Feng Li[1,2,*], Ryan Orban[1,2,†] and Barbara Baker[1,2,*]

[1]*Department of Plant and Microbial Biology, Plant Gene Expression Center, University of California, Berkeley, USA, and*
[2]*United States Department of Agriculture-Agricultural Research Services, 800 Buchanan Street, Albany, CA 94710, USA*

Received 7 December 2011; revised 11 January 2012; accepted 18 January 2012.
*For correspondence (e-mail lifeng@berkeley.edu or bbaker@berkeley.edu).
†Present address: Nutanix Inc. 1735 Technology Dr., Suite 575, San Jose, CA 95110.

SUMMARY

Plant microRNAs (miRNAs) and *trans*-acting small interfering RNAs (tasiRNAs) play important roles in a variety of biological processes. Bioinformatics prediction and small RNA (sRNA) cloning are the most important approaches for identification of miRNAs and tasiRNAs and their targets. However, these approaches are not readily accessible to every researcher. Here we present SoMART, a web server for miRNA/tasiRNA analysis resources and tools, which is designed for researchers who are interested in identifying miRNAs or tasiRNAs that potentially regulate genes of interest. The server includes four sets of tools: 'Slicer detector' for detecting sRNAs targeting input genes, 'dRNA mapper' for detecting degradome (d)RNA products derived from input genes, 'PreMIR detector' for identifying miRNA precursors (MIRs) or tasiRNA precursor (TASs) of input sRNAs, and 'sRNA mapper' for mapping sRNAs onto input genes. We also developed a dRNA-seq protocol to achieve longer dRNA reads for better characterization of miRNA precursors by dRNA mapper. To validate the server function and robustness, we installed sRNA, dRNA and collected genomic DNA or transcriptome databases from Arabidopsis and solanaceous plants, and characterized miR172-mediated regulation of the *APETALA2* gene in potato (*Solanum tuberosum*) and demonstrated conservation of *MIR390*-triggered *TAS3* in tomato (*Solanum lycopersicum*). More importantly, we predicted the existence of *MIR482*-triggered *TAS5* in tomato. We further tested and confirmed the efficiency and accuracy of the server by analyses of 21 validated miRNA targets and 115 miRNA precursors in *Arabidopsis thaliana*. SoMART is available at http://somart.ist.berkeley.edu.

Keywords: web server, miRNA, tasiRNA, degradome RNA, bioinformatics, *Solanum tuberosum*, *Solanum lycopersicum*, *Nicotiana tabacum*, *Arabidopsis thaliana*, technical advance.

INTRODUCTION

micro(mi)RNAs and *trans*-acting small interfering (tasi)RNAs are important regulators of gene expression in plants (Chen, 2008; Rubio-Somoza *et al.*, 2009). They are 20–22 nucleotide (nt) RNAs encoded by RNA polymerase II-dependent genes (Kim *et al.*, 2011). Both miRNAs and tasiRNAs guide Argonaute protein (AGO)-containing complexes or RNA-induced silencing complexes to mediate cleavage of target RNA transcripts or translational inhibition based on sequence complementarity (Brodersen *et al.*, 2008; Mallory *et al.*, 2008). miRNAs are embedded in stem-loop transcripts, or miRNA precursors, and are processed mainly by the enzyme Dicer-like 1 (DCL1). DCL1 cuts the stem-loop at either the base or loop proximal region, releasing a pre-miRNA with a 2 nt 3′ overhang, and then a second DCL1 processing takes place 21 nt proximally from the processed pre-miRNA end, to generate a 21 nt duplex RNA with a 2 nt 3′ overhang at both ends (Dong *et al.*, 2008; Addo-Quaye *et al.*, 2009b; Song *et al.*, 2010). The miRNA/miRNA* duplex is incorporated into the AGO complex, and the passenger strand (miRNA*) is often destroyed (Baumberger and Baulcombe, 2005). However, recent studies have suggested that some miRNA* can also be loaded into AGO proteins and mediate silencing of their targets (Okamura *et al.*, 2008; Zhang *et al.*, 2011). tasiRNAs require more steps to mature.

They are usually encoded by non-structured transcripts, called TAS transcripts, which are targeted by an miRNA (Allen and Howell, 2010). The miRNA directs the AGO complex to cleave the TAS transcript and recruits RNA-dependent RNA polymerase 6 to the cleaved product, which synthesizes double-stranded RNA (dsRNA) using the miRNA cleavage product as template, and then another Dicer-like enzyme, DCL4, is recruited to generate a 21 nt duplex RNA with a 2 nt 3′ overhang from the dsRNA (Allen *et al.*, 2005). Finally, the tasiRNA strand associates with AGO1 and its passenger strand is degraded (Baumberger and Baulcombe, 2005).

Targets of miRNA and tasiRNAs play important roles in many pathways involved in development or responses to the environment (Rubio-Somoza *et al.*, 2009; Covarrubias and Reyes, 2010). Because of their biological importance and the development of next-generation sequencing technology, many conserved miRNAs have been identified in diverse species, and many species-specific miRNAs have been identified (Moxon *et al.*, 2008). However, analytical tools for miRNA-mediated regulation are lagging behind the increasing amount of sRNA data, and are not readily accessible to molecular biologists. Two important questions in analysis of miRNA- and tasiRNA-mediated regulation are target identification and precursor characterization, both of which are technically challenging and require computational expertise. Several online servers have been developed to predict miRNA targets (Rehmsmeier *et al.*, 2004; Zhang, 2005; Maragkakis *et al.*, 2009; Dai and Zhao, 2011). These web tools predict known miRNAs or user-supplied miRNAs that can regulate selected targets, or predict targets for selected miRNAs. A web server has also been implemented for miRNA precursor prediction from input sequences (Jiang *et al.*, 2007). These tools are all based on a computational approach, and thus the results derived from these tools are predictive and lack experimental support.

Here we present a web server that consists of four online tools and four types of databases. The tools include 'Slicer detector', 'dRNA mapper', 'PreMIR detector' and 'sRNA mapper'. The databases include sRNA databases, filtered sRNA (fRNA) databases, degradome RNA (dRNA) databases and DNA databases from three Solanaceae species: *Nicotiana tabacum* (tobacco), *Solanum tuberosum* (potato) and *Solanum lycopersicum* (tomato). Slicer detector and the fRNA databases can be used to predict potential sRNAs targeting an input gene from these species. dRNA mapper and the dRNA databases can be used to validate the predicted sRNA–target regulation by searching for the cleavage product resulting from the predicted cleavage. PreMIR detector and the DNA databases can be used to predict a precursor for the predicted sRNA, and check whether the precursor is an miRNA precursor (MIR) or not. sRNA mapper and dRNA mapper in conjunction with sRNA and dRNA databases can be used to validate the predicted MIR if the sRNA precursor is predicted to be an MIR. These tools and databases can also be used to characterize tasiRNA genes. Databases from other species can also be installed manually to extend the utility of the server. To validate the function and robustness of the server, we identified and characterized potato *stu-MIR172*, which targets the gene *Relative to APETALA2 1* (*RAP1*, FM246879.2). We also identified the tomato *sly-TAS3* precursor and tasiRNAs generated from *sly-TAS3*, and showed that tasiRNAs 5′D7(+) and 5′D8(+) of *sly-TAS3* potentially regulate *Auxin Response Factor 4* (*ARF4*). Furthermore, we used our server to predict the existence of a TAS gene (*sly-TAS5*) that is triggered by *sly-MIR482d*. We further tested and confirmed the efficiency and accuracy of the server by analyses of 21 validated miRNA targets and 115 miRNA precursors in *Arabidopsis thaliana*. These tools and databases can be accessed at http://somart.ist.berkeley.edu.

RESULTS

Identification and characterization of potato *MIR172* targeting *RAP1*

To demonstrate the identification and characterization of an miRNA gene targeting a gene of interest, we used potato *RAP1* (Martin *et al.*, 2009) as an example. To identify by Slicer detector potential sRNAs that can cleave *RAP1* (see Experimental procedures), we pasted the *RAP1* FASTA format sequence in the interface (Figure 1a), selected the fRNAdb STU1 (see Experimental procedures) and then submitted the request. Twenty-three sRNAs were detected (Figure 1b and Data S1.1), and the sRNA with the highest abundance was found to be identical to *miR172* based on a National Center for Biotechnology Information (NCBI) BLAST search (Altschul *et al.*, 1997) and was predicted to cleave *RAP1* after nucleotide 1463 (Figure 1c and Data S1.2). This result is consistent with recent findings in potato (Martin *et al.*, 2009). However, quite a lot of other non-*miR172* sRNAs were also predicted to cleave *RAP1* (Data S1.1 and S1.2).

Recently, high-throughput sequencing-based methods were developed to detect the degradome RNAs that comprise products derived from miRNA- or tasiRNA-mediated cleavage of their targets, providing good experimental evidence for miRNA–target prediction (German *et al.*, 2008; Gregory *et al.*, 2008). We developed a strategy for dRNA sequencing that has a simple protocol and provides long reads for better mapping specificity and miRNA precursor characterization (Figure 2c and Experimental procedures). Using this protocol, we successfully generated three dRNA libraries from tobacco, tomato and potato plants that allow us to detect miRNA- or tasiRNA-mediated cleavage products in these species using dRNA mapper (see Experimental procedures). To test the significance of the predicted sRNAs in *RAP1* regulation, we pasted the *RAP1* FASTA format

Published 2012.
 doi: 10.1111/j.1365-313X.2012.04922.x

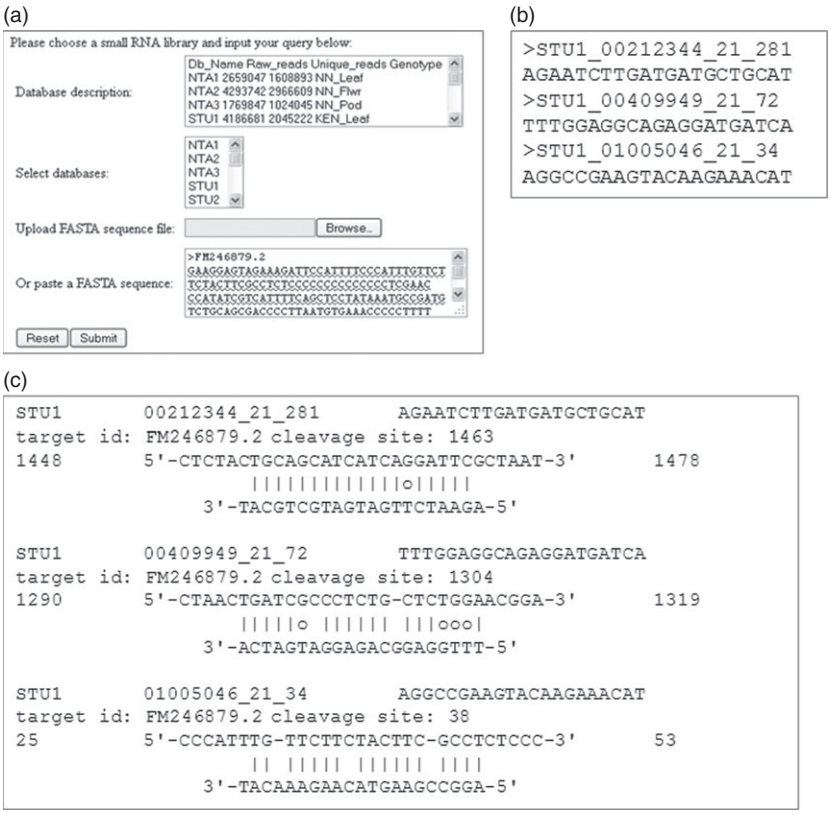

Figure 1. Identification by Slicer detector of sRNAs potentially cleaving *RAP1*.
(a) Interface of Slicer detector.
(b) Partial list of reported slicers in FASTA format that potentially target the potato *AP2* homolog *RAP1* (FM246879). STU1 is the library ID in which the sRNA is identified. 00212344_21_281 is the sRNA ID, and the three numbers in this ID represent the unique numeric ID, length in nucleotides and read numbers of the sRNA, respectively.
(c) Partial report of alignment between *RAP1* and slicers.

sequence into the dRNA mapper interface, selected the dRNAdb StWt and ran the program with default settings (Figure 2a). The results showed that the highest peak of dRNA 5′ end was mapped to position 1464 nt, consistent with the predicted miR172 cleavage site (Figures 1c and 2b), and the other dRNAs detected did not match predicted sRNA cleavage sites. Thus miR172-mediated regulation of *RAP1* is likely to have biological significance. These data show that use of Slicer detector and dRNA mapper together can help to identify relevant sRNA-mediated regulation.

Next we used PreMIR detector to identify the precursor of miR172 in potato. To do this, we pasted the FASTA format of the miR172 sequence into its interface, selected three potato databases and submitted the request (Figure 3a and Experimental procedures). The resulting data showed that two precursor MIRs, *stu-MIR172a* and *stu-MIR172b*, were identified from the EST database, one of which is present in the BAC and NR nucleotide database from NCBI website (Data S3.1). The stem-loop structures of each precursor were visualized using the online UNAFold server (Zuker, 2003) (Figure 3b). We further tested the predicted MIR structure using dRNA mapper, as DCL1 cleavage products of precursor miRNAs were detected for some miRNA genes (Addo-Quaye *et al.*, 2009b). dRNA mapper detected dRNA reads mapped to the position immediately after the 3′ end of mature miRNA on both precursors (Figure 3c and Data S3.2). Among these reads are sequences of 62 and 27 nt long, which can be clearly distinguished from the miRNA contaminations usually found in dRNA-seq (German *et al.*, 2008), indicating that these reads represent real cleavage products and demonstrating that DCL1 processing occurs at predicted positions on both precursors.

According to their mechanism of biogenesis, sRNAs derived from an miRNA precursor should mainly map to the miRNA and miRNA* region, thus we further characterized *stu-MIR172a* and *stu-MIR172b* by sRNA mapper. To analyze their sRNA accumulation pattern, we pasted these two precursor sequences into the sRNA mapper interface, selected the sRNAdb STU1 and then submitted the request with default settings (Figure 4a and Experimental procedures). The results show that the most abundant sRNAs mapped to both precursors were the mature miR172 and the predicted miRNA* (Data S4.1), and the sRNA peaks for these precursors co-localize with the miRNA and miRNA* region (Figure 4b). The statistical analysis produced by sRNA mapper showed that the majority of sRNAs mapped to both precursors are 21 nt in length and contain a 5′ adenosine (Figure 4c), consistent with Arabidopsis miR172 (Park *et al.*, 2002). The sRNA mapping and dRNA mapping data together provide experimental evidence for excision of miRNA/miRNA* duplex from the pre-miRNA structure, which is the primary criterion defining an MIR gene (Meyers *et al.*, 2008).

Published 2012.

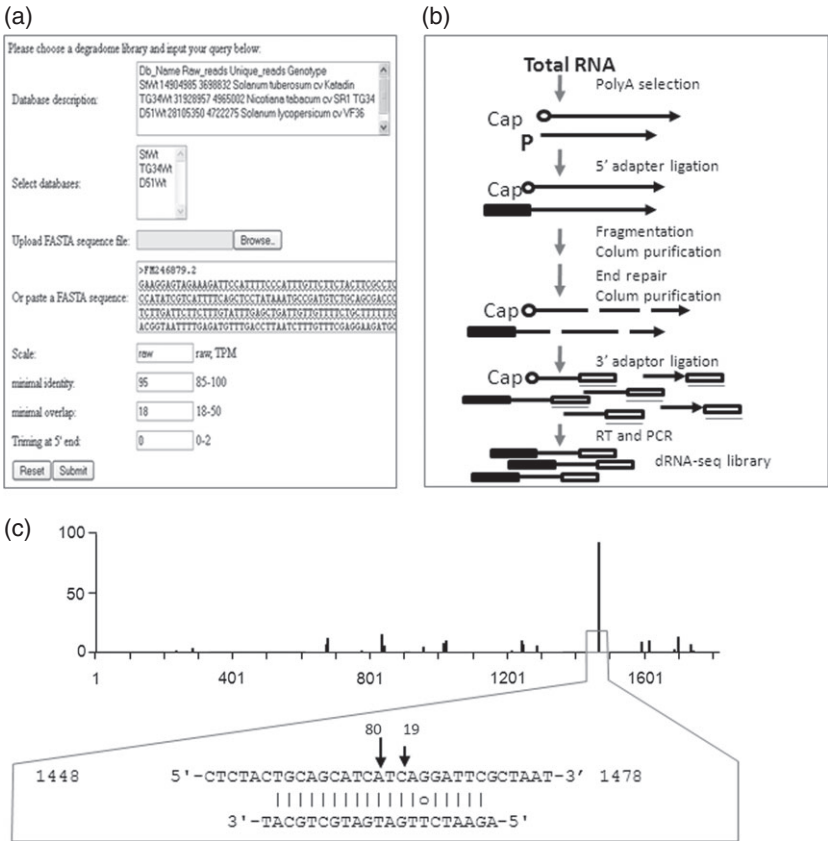

Figure 2. Validation of sRNA–target pairs by dRNA mapper.
(a) Interface of dRNA mapper.
(b) Distribution of dRNA reads on the *RAP1* transcript. The bar graph shows the number of dRNA reads mapped to each position on the transcript (*x* axis) and the sRNA raw read number (*y* axis). The gray box encloses a magnified view of the highest peak region, which overlaps with the slicer-targeted region identified by Slicer detector (Figure 1). The long and short arrows indicate the cleavage sites after 1463 and 1465 nt, to which 80 and 19 dRNA reads mapped, respectively.
(c) Protocol for dRNA-seq library preparation. The open circle represents the mRNA cap structure. The black horizontal line represents mRNA and the arrowhead represents the poly(A) structure. The filled black box represents the 5′ sRNA sequencing adaptor, and the open box represents the 3′ sRNA sequencing adaptor. The gray lines under open boxes represent reverse transcription primers.

Taken together, using our web server, we successfully identified and characterized two loci of *stu-MIR172* that can regulate potato *AP2* homolog *RAP1*.

Identification and characterization of a tomato *TAS3* gene regulating the *ARF4* gene

The above section described the process used to identify and characterize an MIR gene that potentially regulates a known target. It is possible that an sRNA is identified as regulating a target by Slicer detector and dRNA mapper, but the sRNA precursor identified by PreMIR detector may not form a pre-miRNA structure. In this case, it is reasonable to speculate that the sRNA precursor may be a TAS gene. To demonstrate TAS gene characterization, we used the tomato *ARF4* homolog as our gene of interest, which has been shown to be a target of *TAS3* in Arabidopsis (Allen *et al.*, 2005).

The first three steps are similar to identification of miR172 targeting *RAP1*. We used the reported full-length mRNA of tomato *ARF4* (AK319275) (Aoki *et al.*, 2010) and ran it through Slicer detector and dRNA mapper. We identified two sRNAs, D8(+) and D7(+), that were predicted to cleave *ARF4* at two sites (Figure S1A) approximately 200 nt apart, which is similar to the targeting of *ARF4* by *TAS3*-derived tasiRNA in Arabidopsis (Allen *et al.*, 2005). Furthermore, a cleavage site towards the 3′ end was supported by two dRNA reads (Figure S1B). Using PreMIR detector, we identified an EST (DV105041.1) that is capable of producing both sRNAs but cannot fold into a stem-loop structure (Data S5.1). When queried against the TAIR10 transcript database, significant homology was detected between this EST and Arabidopsis *TAS3* (AT3G17185.2) in the region where D8 (+) and D7 (+) reside (e value = 4e−13). Thus we named this EST *sly-TAS3*.

A very important initial indication of a possible TAS gene comes from analyzing its sRNA phasing pattern by sRNA mapper. When we input the *sly-TAS3* FASTA sequence, selected sRNA library SLY2 and ran sRNA mapper, we found that 56% of the sRNAs that mapped to the transcript fell into the same registry, and most were 21 nt long, with a 5′ U residue (Figure 5a,b and Data S5.2), consistent with the features of tasiRNA biogenesis (Allen *et al.*, 2005). Once the sRNA phasing had been confirmed for sRNAs mapped to the putative TAS gene, the next step was to search for a potential miRNA triggering tasiRNA synthesis. For *sly-TAS3*, we identified two abundant sRNAs by Slicer detector, and both were predicted to cleave *sly-TAS3* at position 525 (Figure S2A), which is in-phase with the putative tasiRNAs register (Figure 5b). The more abundant sRNA is identical to Arabidopsis miR390a, while the less abundant sRNA had a G as the third last nucleotide instead of the A in miR390. These results are consistent with the model in Arabidopsis, in which miR390 triggers tasiRNA synthesis on *TAS3* (Allen *et al.*, 2005). Because the majority of dRNA reads are 71 nt in

Published 2012.

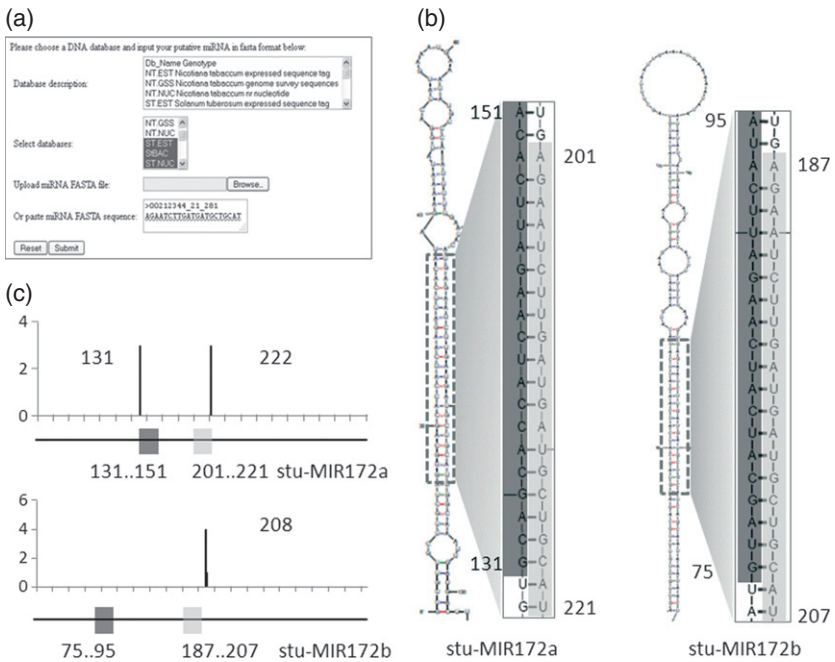

Figure 3. Precursor MIR identification by PreMIR detector and characterization by dRNA mapper.
(a) Interface of PreMIR detector.
(b) Structures of the miRNA precursors generated by the online UNAFold server. The miRNA/miRNA* duplex region is enlarged and shown on the right. The miRNA sequences are highlight in light gray and the miRNA* sequences are highlighted in dark gray, with their coordinates shown next to the end nucleotides.
(c) dRNA maps of *stu-MIR172a* (top) and *stu-MIR172b* (bottom). The numbers in the charts indicate the positions to which dRNAs are mapped. The diagrams under each chart show the position and coordinates of miRNA (light gray) and miRNA* (dark gray) on each precursor predicted by PreMIR detector.

length, we extended *sly-TAS3* to 600 nt using the CoGe web server (Lyons and Freeling, 2008) before searching for dRNA reads matching the predicted cleavage site. Using the extended version of *LeTAS3*, we identified one dRNA read supporting the predicted cleavage using dRNA mapper and the D51 library (Figure 5b).

Finally the tomato *TAS3* trigger was analyzed by PreMIR detector, and two precursors each were identified for *sly-MIR390a*, *b*, *c* and *d* (Data S5.1). They all form stem-loop structures with miR390 in the stem region (Figure S2B and data not shown). However, sRNA mapper only detected the predicted miRNA* for *sly-MIR390b*, and the detected miRNAs are mainly 21 nt RNAs with a 5′ A, as for Arabidopsis miR390 (Figure S2C,D and Data S5.3). These results suggest that *sly-MIR390b* is a bona fide sly-miR390 precursor, while the other three predicted precursors remain to be tested for their ability to produce sly-miR390.

Taken together, the above analyses suggested that, in tomato plants, miR390 could trigger tasiRNA synthesis on *TAS3*, and that the tasiRNA could regulate *ARF4*, indicating that tasiRNA-mediated regulation of *ARF4* is conserved in tomato.

A previously unknown tasiRNA gene in tomato

Plant disease resistance gene families comprised hundreds, even thousands, of members, and are usually targeted by siRNAs and miRNAs (Yi and Richards, 2007; He *et al.*, 2008; Chen *et al.*, 2010; Cuperus *et al.*, 2010). While searching for sRNAs targeting the tomato bacterial spot disease resistance protein 4 gene *Bs4* (cDNA AY438027.1, see also Data S6.2) using Slicer detector, we identified several candidate slicers (Data S6.1). However, no dRNA reads were identified at predicted cleavage sites, so we performed further analysis on the candidate slicer of highest abundance (Figure 6a). A non-MIR precursor locus was detected by PreMIR detector in the tomato genome, and its flanking sequences were extended using CoGe BLAST analysis (Data S6.2 and S6.6). NCBI BLAST analysis identified a stretch of homologous sequence between the precursor and nucleotide-binding-site domain containing (NBS) resistance protein coding sequences (Altschul *et al.*, 1997). sRNA mapper detected abundant siRNAs derived from the precursor, 82% of which fall into the same register (Figure 6b–d and Data S6.3). Thus this precursor is probably a TAS gene, and we named it *sly-TAS5* since there are already four TAS genes discovered in Arabidopsis (Rajagopalan *et al.*, 2006).

Next we searched for a potential miRNA trigger for tasiRNA synthesis from *sly-TAS5* using Slicer detector and dRNA mapper. Among the predicted slicers (Data S6.4), the cleavage directed by a 22 nt slicer was supported by five dRNA reads (Data S6.5), and the cleavage site is in-phase with the major sRNA register (Figure 6d). Querying miRbase with the 22 nt sRNA showed that it is homologous but not identical to *sly-miR482* and *sly-miR482a,b,c*. A putative miRNA precursor of the 22 nt sRNA was identified in the tomato genome by PreMIR detector (Figure 6e and Data S6.6). Thus we named the sRNA *sly-miR482d* and the precursor *sly-MIR482d*. An asymmetric bulge was detected in the miRNA strand of the miRNA/miRNA* duplex, a feature that is important for accumulation of 22 nt rather than 21 nt mature miRNA (Chen *et al.*, 2010; Cuperus *et al.*, 2010). sRNA mapping analysis of *sly-MIR482d* showed that the majority of the sRNAs mapped to the miRNA and miRNA* region on the precursor (Figure S2 and Data S6.7). It has

Published 2012.

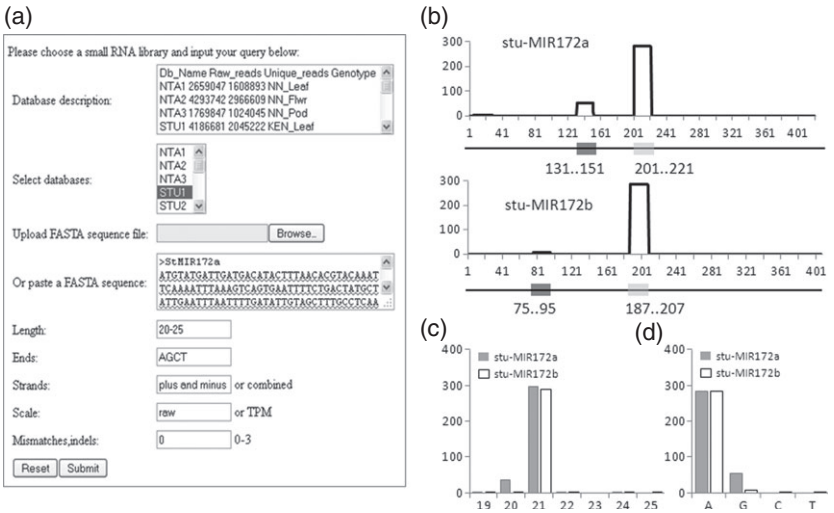

Figure 4. Characterization of precursor MIR by sRNA mapper.
(a) sRNA mapper interface.
(b) sRNA mapping profile of the miR172 precursors. In the chart, the black line represents the number of sRNA reads mapped to the plus strand of the precursor at each nucleotide position; the gray line represents the number of sRNA reads mapped to the minus strand. The *y* axis shows the number of raw reads and the *x* axis shows the nucleotide position. Diagrams below each chart show the precursor structures as in Figure 3.
(c) Length distribution of the sRNAs mapped to the precursor.
(d) Composition of the 5′ end nucleotide of sRNAs mapped to the precursor.

been reported that 22 nt miRNA and tasiRNA have an intrinsic ability to trigger tasiRNA synthesis from the 3′ end of the cleavage product (Chen *et al.*, 2010; Cuperus *et al.*, 2010). Thus these results indicate that *sly-MIR482d* encodes the miRNA that triggers production of phased siRNA from *sly-TAS5*.

These analyses suggested that the tomato genome encodes a previously unknown TAS gene *sly-TAS5* and miRNA gene *sly-MIR482d*, and these two genes interact to generate a tasiRNA that potentially targets resistance gene *Bs4*, demonstrating the ability of our server to identify new miRNA and tasiRNA genes.

Analysis of Arabidopsis miRNAs and targets using SoMART

We further tested the efficiency and accuracy of the predictive function of Slicer detector and PreMIR detector with validated Arabidopsis miRNA targets and miRNAs. We first analyzed validated miRNA targets (Kasschau *et al.*, 2003; Vazquez *et al.*, 2004; Allen *et al.*, 2005) by Slicer detector to determine whether our server can retrieve the correct targeting miRNAs. The corresponding miRNAs were detected for all 21 targets tested, and dRNA reads supporting the predicted cleavage sites were detected for 20 of them (Data S7). AT2G33770.1 was shown to have five target sites for miRNA399, four of which were cleavable. Our server detected all four cleavable sites and retrieved dRNAs supporting the cleavage (Data S7). When we analyzed these validated targets using the same sRNA and dRNA databases and CleaveLand software (Addo-Quaye *et al.*, 2009a), miRNAs for 19 of the 21 targets were identified (Data S7). These results demonstrate that Slicer detector can accurately identify miRNA regulators of input targets, and, when combined with dRNA mapper, the server predicts targets with an accuracy rate of 95%, similar to the existing tools. Next we used the mature sequences listed in miRbase ranging from miR156 to miR447 as input to test the ability of PreMIR detector to detect their precursor sequences. miRbase recorded 124 precursors for these mature miRNAs. PreMIR detector detected the genetic loci for all pre-miRNA, and 111 of them were predicted to be miRNAs (Data S8). Further examination of 13 loci that were not predicted to be miRNAs revealed that six of them (miR413, 414, 415, 418, 419, 420 and 426) had too many bulges or mismatches in the predicted miRNA/miRNA* duplex region, and were annotated 'status in question' on miRbase. The other three (miR401, 404 and 406) had similar problems to the aforementioned six, although they were not annotated 'status in question'. Thus PreMIR detector detected 111 of the 115 valid precursors, an accuracy of 96.5%. Taken together, these analyses confirm that the server is robust enough to identify most, if not all, miRNA-mediated regulation and miRNA precursors.

DISCUSSION

As miRNAs and tasiRNAs are recognized as important players in more and more biological processes, identification of miRNA/tasiRNA–target pairs has become critical for probing the regulatory network of many pathways. Solanaceous species include many important crops, and are important model species for a number of biological processes of agricultural importance, including plant–pathogen interactions, fruit ripening and tuber formation. Recent studies showed that some of these processes are regulated by miRNAs (Moxon *et al.*, 2008; Martin *et al.*, 2009; Xie *et al.*, 2011). Here we present a web server that consists of four analysis tools linked to various Solanaceae databases. Researchers can use these tools and databases to identify and characterize miRNA and tasiRNA components in the pathways of three Solanaceae species. To achieve this goal, we recommend the following steps (Figure 7): (i) start with any gene of interest and use Slicer detector to identify potential sRNA regulators; (ii) validate the predicted regulation by dRNA mapper to search

Published 2012.

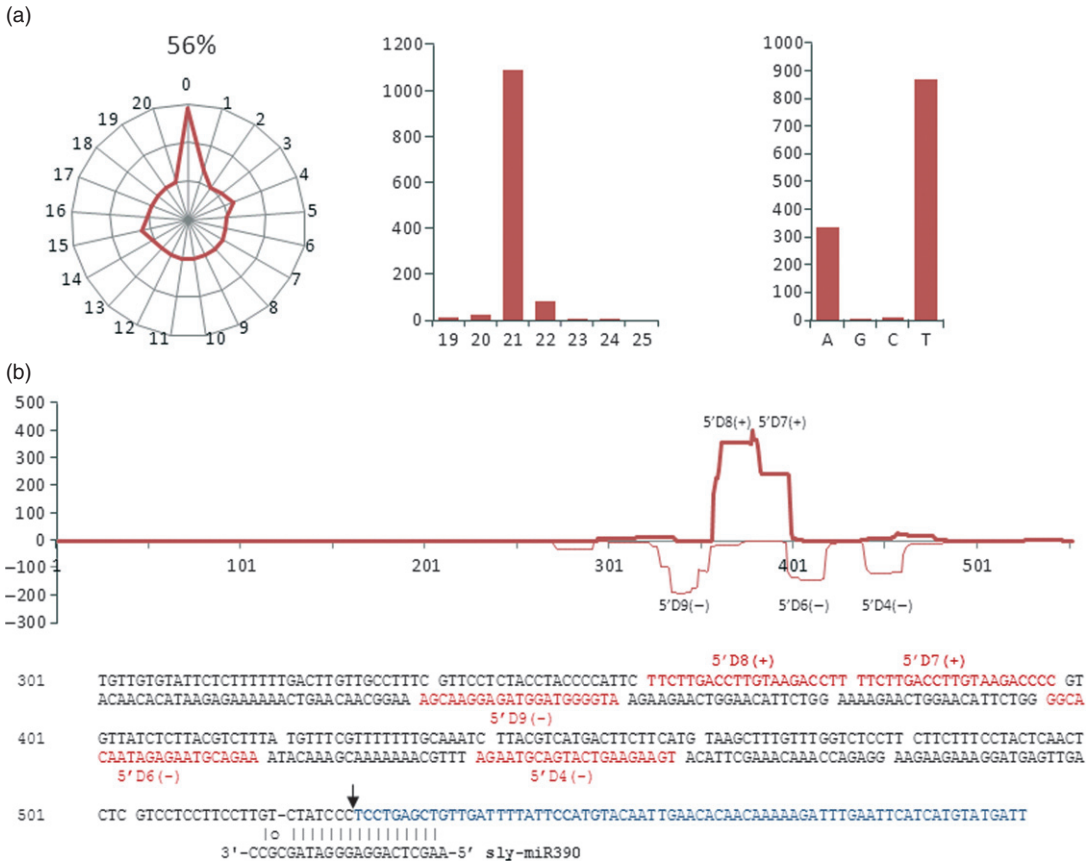

Figure 5. Characterization of *sly-TAS3*.
(a) Left: register of sRNAs mapped to *sly-TAS3*, generated by Excel using sRNA phasing data generated by sRNA mapper. For each sRNA, the register *R* was calculated using the formula $R = P - 21N$, where *P* is the coordinate of the 5′ end of the sRNA if the sRNA matches the plus strand, or 2 plus the coordinate of the 3′ end of the sRNA if it matches the minus strand, and *N* is the maximal integer that allows *R* to be greater than 0. The total count of sRNA in each register was divided by the total number of sRNAs mapped and plotted against the register number using a radar chart. The inner circle represents 0%, the outside circle represents 60%, and the middle circle represents 30%. Middle: length distribution of the sRNAs mapped to *sly-TAS3*. Right: 5′ end composition of sRNAs mapped to *sly-TAS3*.
(b) Top: sRNA mapping profile of *sly-TAS3*. Bottom: partial sequence of *sly-TAS3* with cloned tasiRNAs highlighted in red. The cleavage site supported by dRNA read is indicated by an arrow. The nomenclature of the tasiRNAs was adopted from Allen *et al.* (2005), and their positions are marked on the map. The blue text indicates where dRNA read is mapped.

for evidence of sRNA-mediated cleavage at predicted sites; (iii) identify the precursor for the sRNA by PreMIR detector if a regulatory sRNA is identified from the first two steps; (iv) characterize the pre-miRNA by sRNA mapper and dRNA mapper. If a possible TAS precursor is found in step 3, start with the putative TAS precursor and repeat steps 1–4 to identify and characterize the triggering miRNA. As proof of concept, we successfully identified and characterized *stu-MIR172*, which regulates the potato *AP2* homolog *RAP1*, and also showed that *TAS3* tasiRNA-mediated regulation of *ARF4* is likely in tomato. More importantly, our analysis predicted the existence of a TAS gene (*sly-TAS5*) that is triggered by *sly-MIR482d* in tomato. In a related study, we successfully identified and characterized ten families of previously unknown miRNAs targeting various classes of functional resistance genes in three solanaceous species using our server (Li *et al.*, 2012). The robustness and accuracy of the server's predictive power were further tested and confirmed with 21 validated Arabidopsis miRNA targets and 115 Arabidopsis miRNA precursors.

We also developed a dRNA-seq protocol that is simple and produces long dRNA reads. The longer read length of dRNA provides better specificity in sequence mapping. It is of particular importance in miRNA precursor characterization because it can distinguish DCL1 cleavage products from the miRNA or miRNA* contamination that is often associated with dRNA-seq. In contrast, the restriction-based dRNA-seq protocol can only generate approximately 20 nt reads, and thus cannot distinguish between contamination and real cleavage products (German *et al.*, 2008).

These tools can be used to characterize both known and unknown miRNA- and tasiRNA-mediated regulation of genes of interest. However, due to the limited number of datasets, some sRNA-mediated regulation may be missed because of tissue-specific or induced expression of certain sRNAs. With the capacity to install additional sRNA and

Published 2012.

dRNA databases and upgrades in genomic sequences, our web server is a useful platform for studying miRNA- and tasiRNA-mediated regulation in Solanaceae and other species in general. It is also useful to implement recently proposed criteria for miRNA annotation in Solanaceae species (Meyers *et al.*, 2008). With its simple user-friendly interface and flexible data processing, this server can help biologists identify targets in order to focus on the biological function of miRNA/tasiRNA-mediated gene regulation, rather than spending a lot of time cloning sRNAs or dRNAs and mastering computational techniques.

EXPERIMENTAL PROCEDURES

dRNA-seq

A protocol for sequencing dRNA was developed to obtain long reads of dRNA for more specific alignment to a gene of interest (Figure 2c). In brief, poly(A) RNA is isolated from total RNA using oligo(dT) magnetic beads (Invitrogen, http://www.invitrogen.com/) and then a 5′ sRNA sequencing adapter is ligated. The ligation product is purified using oligo(dT) beads, and fragmented in RNA fragmentation buffer (Ambion, www.invitrogen.com). The fragmented RNA is purified using an RNeasy micro kit (Qiagen, http://www.qiagen.com/) and treated with Antarctic phosphatase (NEB, www.neb.com) to repair the 3′ end, and then the 3′ sRNA sequencing adapter is ligated. The dRNA-seq library is obtained after RT-PCR and gel purification of 150–300 nt DNA fragments. The library is sequenced using an Illumina Genome Analyzer (Illumina, www.illumina.com) with sRNA-seq primers.

Database description

The web server uses four sets of databases for MIR and TAS gene identification and characterization. Thirteen sRNA databases (sRNAdb) were generated and linked to sRNA mapper for mapping sRNA onto input genes. To generate these sRNA databases, sRNA libraries were prepared according to the Illumina sRNA-seq protocol, and sequenced using an Illumina Genome Analyzer. Thirteen filtered sRNA databases (fRNAdb) were generated and linked to Slicer detector to detect potential slicers of input genes. Each fRNAdb is a subset of sRNAs with five or more raw reads in the corresponding sRNAdb. Three dRNA databases (dRNAdb) were generated and linked to dRNA mapper for detecting cleavage products derived from input genes. To produce these dRNA databases, dRNA libraries were prepared using the protocol

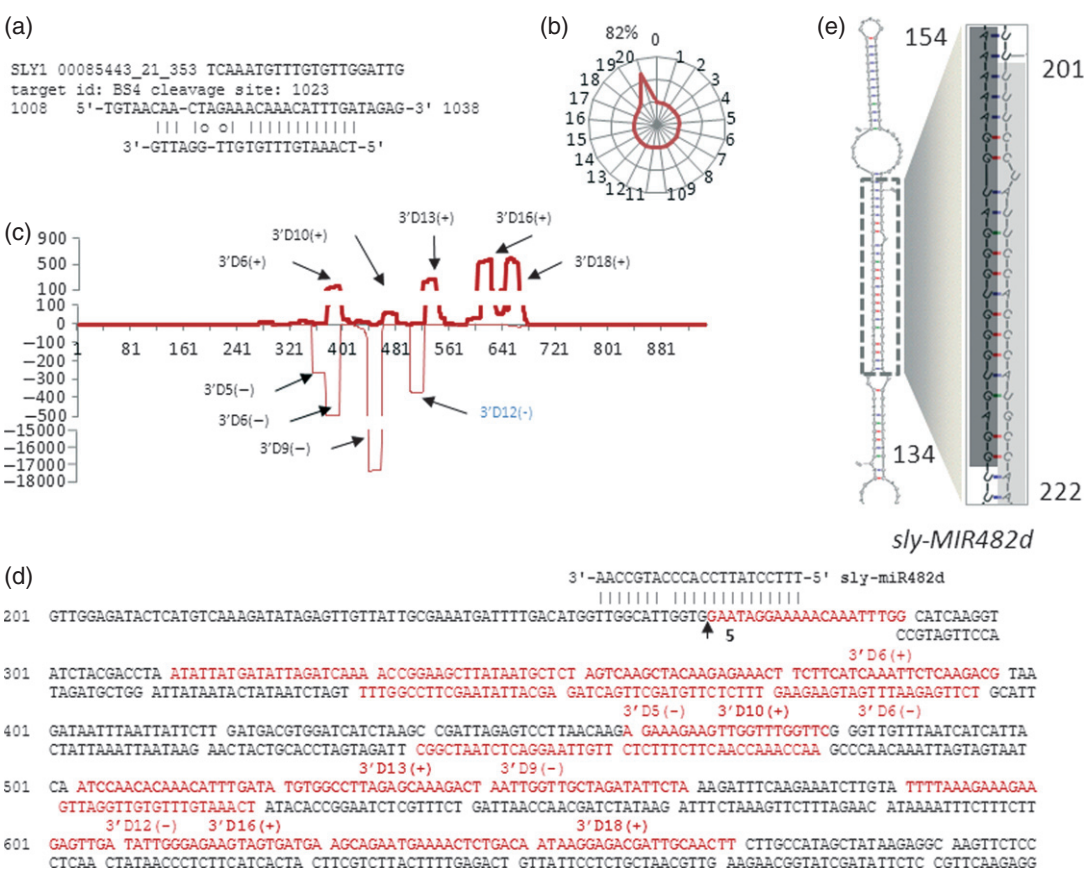

Figure 6. Identification of a previously unknown tasiRNA gene, *sly-TAS5*, in tomato.
(a) Identification of a 21 nt sRNA potentially targeting the bacterial spot resistance gene *Bs4* using Slicer detector.
(b) Phasing pattern of sRNA mapped to *sly-TAS5*, generated as described in Figure 5(a). The outside circle represents 120% and the middle circle represents 60%.
(c) sRNA map of *sly-TAS5* generated by Excel based on data from sRNA mapper.
(d) Partial sequences of *sly-TAS5* labeled as in Figure 5(b).
(e) Structure of *sly-MIR482d*, generated as described in Figure 3(b).

Published 2012.

described in Figure 2(c) and sequenced using an Illumina Genome Analyzer. Ten genomic DNA or EST databases were downloaded from public databases and linked to PreMIR detector for miRNA and tasiRNA precursor identification. The EST, Nucleotide and GSS databases were batch-downloaded from the NCBI EST, NR nucleotide and genome survey sequence databases, respectively. Potato BAC sequences were downloaded from NCBI using the species name '*Solanum tuberosum*' and the keyword 'BAC'. The *Solanum tuberosum* Phureja genome and *Solanum lycopersicum* genome were downloaded from the Potato Genome Sequencing Consortium database (Xu *et al.*, 2011, http://potatogenomics.plantbiology.msu.edu/index.html) and the International Tomato Genome Sequencing Consortium database (http://solgenomics.net/organism/Solanum_lycopersicum/genome), respectively. A summary of all databases used in this study is provided in Table S1 (Whitham *et al.*, 1994, 1996; German *et al.*, 2008; Kuang *et al.*, 2009; Mahalingam and Meyers, 2010; Zheng *et al.*, 2010; Xu *et al.*, 2011).

Slicer detector

Slicer detector detects sRNAs that can potentially cleave an input transcript sequence. It is linked to the fRNAdb, which contain sRNA sequences in FASTA format with a unique ID (Figure 1b). As input, it uses a FASTA format sequence, for example *RAP1*, without special characters in the sequence ID and an fRNAdb selected from the pull-down menu (Figure 1a). The background script uses BLASTN (Altschul *et al.*, 1997) to query the input FASTA sequence against the selected fRNAdb with very low stringency (word_size = 7, evalue = 10 000) and retrieves all matching sRNAs. For each alignment, sequences of approximately 30 nt covering the aligned input region are retrieved and aligned to the matching sRNAs using a modified Smith–Waterman algorithm to allow G to pair with T and T to pair with G. Finally, the new alignment is examined for cleavability based on criteria proposed previously (Allen *et al.*, 2005; Schwab *et al.*, 2005), and cleavable slicer–target pairs are reported in two files. One shows the FASTA format sequences of potential slicers for the input (Data S1.1). The other shows the alignments between the input and each slicer, with the predicted cleavage site on the target (Data S1.2).

dRNA mapper

dRNA mapper provides experimental evidence for sRNA-directed cleavage by mapping dRNA reads to an input target sequence. As input, it uses a FASTA format sequence of the gene of interest and a selection of dRNAdb from the pull-down menu. The background script uses BLASTN (Altschul *et al.*, 1997) to query the input FASTA sequence against the selected dRNAdb(s) using default BLASTN settings, and reports the matching dRNA reads in a table (Data S2.1). It also creates a dRNA read map of the input transcript at single-nucleotide resolution, and reports the data in a three-column table (Data S2.2). Using this table, a chart can be created in Excel to show the distribution of the dRNA reads on the transcript.

PreMIR detector

PreMIR detector identifies precursor sequences of input sRNAs and checks whether the precursor sequence can fold into the pre-miRNA structure. As input, it uses a FASTA format sRNA sequence and a selection of DNA database(s). The background program uses BLASTN (Altschul *et al.*, 1997) to query the input sRNA sequence(s) against the selected DNA databases to find perfect matching DNA sequences, retrieves up to 200 nt flanking sequences, and reports the DNA sequence as sRNA precursor. Next it uses the UNAFold program (Zuker, 2003) to fold the precursor and predict the coding arms for the miRNA, and coordinates for the miRNA* if the precursor can fold into a stem-loop structure with the sRNA in the stem. If no sequence can be found matching the sRNA, it reports 'no genetic locus detected'. If a perfect matching sequence is found that does not fold into a pre-miRNA structure, no miRNA coding arm or miRNA* coordinates are predicted. Multiple sRNA FASTA sequences and multiple DNA databases can be used concurrently.

sRNA mapper

sRNA mapper maps sRNAs from selected sRNAdb(s) to an input gene. As input, it uses a FASTA format sequence for a gene of interest and selected sRNAdb. The background program uses BLASTN (Altschul *et al.*, 1997) to query the input gene sequence against the selected sRNAdb(s) to find matching sRNAs (with word_size 11 and num_alignment 8000 settings). Four report files are generated. The first lists the sRNA reads mapped to the gene of interest, with the number of mismatches or indels set by the user. The second shows the phasing pattern of sRNAs mapped to the input sequence, which is useful for TAS gene characterization. The third shows the statistics for the length distribution and 5′ end composition of sRNAs. The fourth reports an sRNA mapping profile in a table similar to that of dRNA mapper. There are options for the user to plot a subset of sRNAs with selected size and 5′ ends, in raw reads or transcript per million (TPM) scale, to show sRNA strandedness or not for the mapping profile. These data can be downloaded and processed by Excel to generate charts showing the

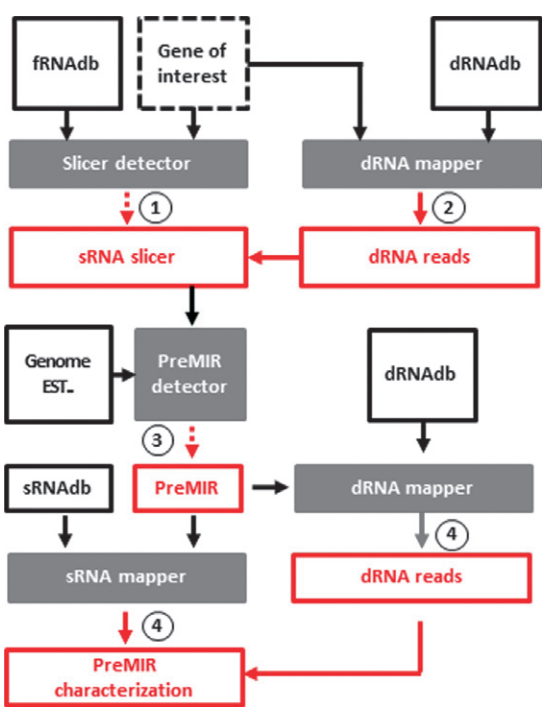

Figure 7. Recommended pipeline for identification and characterization of MIR genes targeting genes of interest.
The open box with a dashed black line represents the gene of interest as the starting point of analysis. Filled gray boxes represent analysis tools. Open boxes with a solid black line represent databases linked to various tools. Black arrows represent data input. Dashed red arrows represent bioinformatics predictions from the web server. Solid red arrows represent experimental data retrieved from the web server that provide supporting evidence for predictions. The numbers show the steps of the recommended pipeline.

Published 2012.

distribution of sRNAs on the input sequence, the length distribution and the 5′ end composition of mapped sRNAs.

Implementation of the web server

In order to facilitate use of these tools, a web interface was created based on the Common Gateway Interface (CGI, http://www.w3.org/CGI/) protocol, which provides a mechanism for the web server to generate pages to send to the user and for the user to specify data and options to the script. Written entirely in Perl, the interface used the Comprehensive Perl Archive Network (CPAN) packages CGI.pm, CGI::FormBuilder, and CGI::Session (http://www.cpan.org/) to produce the necessary HTML and JavaScript, while the open-source HTTP web server Apache (http://httpd.apache.org/) was used to mediate transactions between the script and the user. Query options were collected by form submission together with the FASTA sequence to be analyzed, either by pasting the sequence directly or uploading a file. Once submitted, the analysis script is invoked via a child process, and the results are presented directly to the user for display, download or further analysis. A unique identifier is generated for each user and stored in the form of a cookie in the user's browser, allowing multiple concurrent users. The web server uses Ubuntu 10.04 LTS Lucid Lynx, a free and open-source Linux operating system that is available to the public, running on a quad-core hyper-threaded Nehalem Intel i7 920 processor with 12 GB of RAM.

ACKNOWLEDGEMENTS

The authors are grateful to Professor Sheila McCormick, Department of plant and microbial biology of University of California Berkeley for critical reading of the manuscript and providing valuable comments. This work was supported by the National Science Foundation Plant Genome Research Program Grant (DBI-0218166) and the United States Department of Agriculture (CRIS 5335-22000-007-00D). The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. *TAS3* tasiRNA-mediated *ARF4* regulation.

Figure S2. Characterization of *sly-MIR390* and *sly-MIR482d*.

Table S1. Summary of databases used by the web server.

Data S1. Slicer detector result for *RAP1* (FM246879.2).

Data S2. dRNA mapper result for *RAP1* (FM246879.2).

Data S3. Characterization of *stu-MIR172* by PreMIR detector and dRNA mapper.

Data S4. Characterization of *stu-MIR172* by sRNA mapper.

Data S5. Characterization of *sly-TAS3* and *sly-MIR390*.

Data S6. Characterization of *sly-TAS5* and *sly-MIR482d*.

Data S7. Validation of Slicer detector with published Arabidopsis miRNA targets.

Data S8. Prediction of Arabidopsis miRNA precursors by PreMIR detector.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES

Addo-Quaye, C., Miller, W. and Axtell, M.J. (2009a) CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics*, **25**, 130–131.

Addo-Quaye, C., Snyder, J.A., Park, Y.B., Li, Y.F., Sunkar, R. and Axtell, M.J. (2009b) Sliced microRNA targets and precise loop-first processing of *MIR319* hairpins revealed by analysis of the *Physcomitrella patens* degradome. *RNA*, **15**, 2112–2121.

Allen, E. and Howell, M.D. (2010) miRNAs in the biogenesis of *trans*-acting siRNAs in higher plants. *Semin. Cell Dev. Biol.* **21**, 798–804.

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.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Aoki, K., Yano, K., Suzuki, A. *et al.* (2010) Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the Solanaceae genomics. *BMC Genomics*, **11**, 210.

Baumberger, N. and Baulcombe, D.C. (2005) Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 11928–11933.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L. and Voinnet, O. (2008) Widespread translational inhibition by plant miRNAs and siRNAs. *Science*, **320**, 1185–1190.

Chen, X. (2008) MicroRNA metabolism in plants. *Curr. Top. Microbiol. Immunol.* **320**, 117–136.

Chen, H.M., Chen, L.T., Patel, K., Li, Y.H., Baulcombe, D.C. and Wu, S.H. (2010) 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proc. Natl Acad. Sci. USA*, **107**, 15269–15274.

Covarrubias, A.A. and Reyes, J.L. (2010) Post-transcriptional gene regulation of salinity and drought responses by plant microRNAs. *Plant Cell Environ.* **33**, 481–489.

Cuperus, J.T., Carbonell, A., Fahlgren, N. *et al.* (2010) Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. *Nat. Struct. Mol. Biol.* **17**, 997–1003.

Dai, X. and Zhao, P.X. (2011) psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* **39**, W155–W159.

Dong, Z., Han, M.H. and Fedoroff, N. (2008) The RNA-binding proteins HYL1 and SE promote accurate *in vitro* processing of pri-miRNA by DCL1. *Proc. Natl Acad. Sci. USA*, **105**, 9970–9975.

German, M.A., Pillay, M., Jeong, D.H. *et al.* (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.* **26**, 941–946.

Gregory, B.D., O'Malley, R.C., Lister, R. *et al.* (2008) A link between RNA metabolism and silencing affecting Arabidopsis development. *Dev. Cell.* **14**, 854–866.

He, X.F., Fang, Y.Y., Feng, L. and Guo, H.S. (2008) Characterization of conserved and novel microRNAs and their targets, including a TuMV-induced TIR-NBS-LRR class R gene-derived novel miRNA in Brassica. *FEBS Lett.* **582**, 2445–2452.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. and Lu, Z. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* **35**, W339–W344.

Kasschau, K.D., Xie, Z., Allen, E., Llave, C., Chapman, E.J., Krizan, K.A. and Carrington, J.C. (2003) P1/HC-Pro, a viral suppressor of RNA silencing, interferes with Arabidopsis development and miRNA function. *Dev. Cell.* **4**, 205–217.

Kim, Y.J., Zheng, B., Yu, Y., Won, S.Y., Mo, B. and Chen, X. (2011) The role of Mediator in small and long noncoding RNA production in *Arabidopsis thaliana*. *EMBO J.* **30**, 814–822.

Kuang, H., Padmanabhan, C., Li, F. *et al.* (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res.* **19**, 42–56.

Li, F., Pignatta, D., Bendix, C. *et al.* (2012) MicroRNA regulation of plant innate immune receptors. *Proc. Natl Acad. Sci. USA*, **109**(5), 1790–1795. doi: pnas.1118282109.

Lyons, E. and Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53**, 661–673.

Mahalingam, G. and Meyers, B.C. (2010) Computational methods for comparative analysis of plant small RNAs. *Methods Mol. Biol.* **592**, 163–181.

Mallory, A.C., Elmayan, T. and Vaucheret, H. (2008) MicroRNA maturation and action – the expanding roles of ARGONAUTEs. *Curr. Opin. Plant Biol.* **11**, 560–566.

Published 2012.

Maragkakis, M., Reczko, M., Simossis, V.A. *et al.* (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.* **37**, W273–W276.

Martin, A., Adam, H., Diaz-Mendoza, M., Zurczak, M., Gonzalez-Schain, N.D. and Suarez-Lopez, P. (2009) Graft-transmissible induction of potato tuberization by the microRNA miR172. *Development*, **136**, 2873–2881.

Meyers, B.C., Axtell, M.J., Bartel, B. *et al.* (2008) Criteria for annotation of plant microRNAs. *Plant Cell*, **20**, 3186–3190.

Moxon, S., Jing, R., Szittya, G., Schwach, F., Rusholme Pilcher, R.L., Moulton, V. and Dalmay, T. (2008) Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Res.* **18**, 1602–1609.

Okamura, K., Phillips, M.D., Tyler, D.M., Duan, H., Chou, Y.T. and Lai, E.C. (2008) The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. *Nat. Struct. Mol. Biol.* **15**, 354–363.

Park, W., Li, J., Song, R., Messing, J. and Chen, X. (2002) CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12**, 1484–1495.

Rajagopalan, R., Vaucheret, H., Trejo, J. and Bartel, D.P. (2006) A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.* **20**, 3407–3425.

Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Rubio-Somoza, I., Cuperus, J.T., Weigel, D. and Carrington, J.C. (2009) Regulation and functional specialization of small RNA-target nodes during plant development. *Curr. Opin. Plant Biol.* **12**, 622–627.

Schwab, R., Palatnik, J.F., Riester, M., Schommer, C., Schmid, M. and Weigel, D. (2005) Specific effects of microRNAs on the plant transcriptome. *Dev. Cell*, **8**, 517–527.

Song, L., Axtell, M.J. and Fedoroff, N.V. (2010) RNA secondary structural determinants of miRNA precursor processing in Arabidopsis. *Curr. Biol.* **20**, 37–41.

Vazquez, F., Gasciolli, V., Crete, P. and Vaucheret, H. (2004) The nuclear dsRNA binding protein HYL1 is required for microRNA accumulation and plant development, but not posttranscriptional transgene silencing. *Curr. Biol.* **14**, 346–351.

Whitham, S., Dinesh-Kumar, S.P., Choi, D., Hehl, R., Corr, C. and Baker, B. (1994) The product of the tobacco mosaic virus resistance gene *N*: similarity to Toll and the interleukin-1 receptor. *Cell*, **78**, 1101–1115.

Whitham, S., McCormick, S. and Baker, B. (1996) The *N* gene of tobacco confers resistance to tobacco mosaic virus in transgenic tomato. *Proc. Natl Acad. Sci. USA*, **93**, 8776–8781.

Xie, F., Frazier, T.P. and Zhang, B. (2011) Identification, characterization and expression analysis of MicroRNAs and their targets in the potato (*Solanum tuberosum*). *Gene*, **473**, 8–22.

Xu, X., Pan, S., Cheng, S. *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

Yi, H. and Richards, E.J. (2007) A cluster of disease resistance genes in Arabidopsis is coordinately regulated by transcriptional activation and RNA silencing. *Plant Cell*, **19**, 2929–2939.

Zhang, Y. (2005) miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res.* **33**, W701–W704.

Zhang, X., Zhao, H., Gao, S. *et al.* (2011) *Arabidopsis* Argonaute 2 regulates innate immunity via miRNA393*-mediated silencing of a Golgi-localized SNARE gene, *MEMB12*. *Mol. Cell*, **42**, 356–366.

Zheng, Q., Ryvkin, P., Li, F. *et al.* (2010) Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.* **6**(9), e1001141.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415.

Published 2012.
 doi: 10.1111/j.1365-313X.2012.04922.x