Uncovering DCL1-dependent small RNA loci on plant genomes: a structure-based approach

Yijun Meng1,*, Chaogang Shao2, Huizhong Wang1 and Ming Chen3,*

1 College of Life and Environmental Sciences, Hangzhou Normal University, Hangzhou 310036, PR China
2 College of Life Sciences, Huzhou Teachers College, Huzhou 313000, PR China
3 Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, PR China

* To whom correspondence should be addressed. E-mail: mengyijun@zju.edu.cn or mchen@zju.edu.cn

Received 30 August 2013; Revised 24 October 2013; Accepted 8 November 2013

Abstract

In plants, Dicer-like 1 (DCL1)-mediated two-step cleavages are essential for the processing of microRNA (miRNA) gene products. Interestingly, DCL1 has been indicated to be involved in the production of many small RNAs (sRNAs) that cannot be classified as canonical miRNAs. However, genomic and functional information on the non-miRNA, DCL1-dependent sRNAs is still limited. Here, we propose a secondary structure-based approach for identification of the precursors containing novel DCL1-dependent sRNA loci. To demonstrate the utility of the workflow: first, 5898 DCL1-dependent sRNAs of 20–24 nucleotides were identified from the sRNA high-throughput sequencing data sets prepared from rice DCL1 RNA interference transgenic lines. Those perfectly mapped to the rice pre-miRNAs (precur- sors) were removed. The remaining 5795 sRNAs were then mapped onto the rice genome, obtaining 30 902 perfectly matched loci belonging to 2310 sRNAs. A total of 4631 clusters of sRNA loci were defined for secondary structure prediction by using RNAfold. The prediction results generated by two algorithms, namely MFE (minimum free energy) and centroid, were manually compared to identify the conserved long-stem structures containing DCL1-dependent sRNA loci. For the purpose of a case study, a portion of the prediction results was screened manually. As a result, 60 clusters displayed great potential for forming featured long-stem structures for the generation of DCL1-dependent sRNAs. Together, the results indicate that the proposed workflow is applicable for the identification of novel DCL1-dependent sRNA loci on plant genomes.

Key words: Cluster, Dicer-like 1 (DCL1), high-throughput sequencing, microRNA (miRNA), precursor microRNAs (pre-miRNAs), secondary structure, small RNA (sRNA).

Introduction

The essential regulatory roles of plant microRNAs (miRNAs) have been widely recognized during the last 10 years (Jones-Rhoades et al., 2006; Voinnet, 2009). These ~21 nucleotide (nt) small RNAs (sRNAs) are processed from stem–loop structured precursors through Dicer-like 1 (DCL1)-mediated two-step cleavages, which is considered to be the featured procedure for the biogenesis of plant miRNAs. With the advent of high-throughput sequencing (HTS), the universe of plant sRNAs has been greatly expanded. Several new sRNA species, such as nat-siRNAs (natural antisense small interfering RNAs) (Borsani et al., 2005), ta-siRNAs (trans-acting small interfering RNAs) (Vazquez et al., 2004; Allen et al., 2005; Williams et al., 2005), and ra-siRNAs (repeat-associated small interfering RNAs) (Herr et al., 2005; Onodera et al.,
of the proposed workflow, a case study was carried out in rice. As a result, many long-stem structured precursors encoding DCL1-dependent sRNAs were identified, which could serve as the basis for in-depth experimental analyses. On the other hand, the results indicate that the structure-based approach was applicable for the identification of DCL1-dependent sRNA loci on plant genomes.

Secondary structure-based approach to identify DCL1-dependent sRNA loci on plant genomes

The sRNA HTS data generated from the DCL1 RNA interference (RNAi) transgenic lines or dcl1 mutants could be utilized, treating the data from the wild-type plants as the control. In order to allow cross-library comparison, the normalized read count (in RPM, reads per million) of an sRNA from a specific data set was calculated by dividing the raw count of this sRNA by the total counts of the data set, and then multiplied by 10^6. In the workflow (Fig. 1A), we apply the following criteria to extract the DCL1-dependent sRNAs: for a group of sRNA HTS data (ideally generated by one experiment), the normalized accumulation level

![Fig. 1. Structure-based workflow for the identification of the potential long-stem structured precursors encoding Dicer-like 1-dependent small RNAs. The workflow is shown on the left-hand side of (A) and (B), and the results of the case study are shown on the right-hand side. (This figure is available in colour at JXB online.)](https://example.com/fig1.png)
of an sRNA should be ≥3 RPM (adjustable parameter) in at least one of the control data sets, and should be three times or more (adjustable parameter) than its level in the DCL1 RNAi or dcl1 mutant data set. The DCL1-dependent sRNAs of 20–24 nt (adjustable parameter) are retained, considering the fact that most of the plant miRNAs fall within this length range [e.g. 337 out of 337 miRBase-registered (Griffiths-Jones et al., 2008) miRNAs in Arabidopsis and 709 out of 713 registered miRNAs in rice fall within this range].

As introduced above, the miRNA precursors are processed in a DCL1-dependent pathway in plants. Thus, to identify novel genomic loci, the obtained DCL1-dependent sRNAs are mapped to the miRBase-registered pre-miRNAs (precursor microRNAs) of the plant species analysed. The sRNAs perfectly mapped onto the pre-miRNAs are removed (Fig. 1A).

Search for clusters of DCL1-dependent sRNAs on plant genomes

All of the DCL1-dependent sRNAs identified above are mapped to the plant genome, and the perfectly matched loci are reserved.

To search for the potential precursors generating DCL1-dependent sRNAs on the plant genome, sRNA clusters are defined. A cluster is constituted by two or more DCL1-dependent sRNA loci resided within a genomic region of ≤500 nt (adjustable parameter) (Fig. 1A).

Secondary structure-based identification of long-stem regions encoding DCL1-dependent sRNAs on the precursors

According to the strand bias of the sRNA loci within a defined cluster, a specific genomic strand is selected for secondary structure prediction by using RNAfold (Hofacker, 2003). Two distinct algorithms, minimum free energy (MFE) and centroid, are employed by RNAfold during secondary structure prediction. The MFE-based structure is predicted using a loop-based energy model and the dynamic programming algorithm (Zuker and Stiegler, 1981). The predicted secondary structure possesses a minimum free energy. The centroid structure of an RNA sequence is the secondary structure with a minimal base pairing distance to all other secondary structures in the Boltzmann ensemble (Ding et al., 2005). Based on the biogenesis model of the plant miRNAs, the highly complementary stem regions of the hairpin structured precursors are the ideal substrates for DCL1-mediated processing. In this regard, we set out to perform a manual screening for the conserved long-stem regions encompassing DCL1-dependent sRNA loci through structure comparison between MFE- and centroid-based prediction results (Fig. 1B).

To find evidence for expression of the potential precursors encoding DCL1-dependent sRNAs, ESTs (expressed sequence tags), RNA-seq (RNA sequencing), and microarray data could be utilized (Fig. 1B).

Case study

To demonstrate the applicability of the workflow proposed above, we carried out a comprehensive search for the potential DCL1-dependent sRNA loci on the rice genome. First, three groups of sRNA HTS data sets were retrieved from public databases. The first group contains four data sets (P1dc1I, P1dc1D, P2dc1I, and P2dc1D) generated from the panicles of the rice DCL1 RNAi transgenic lines, and 15 data sets (PC1I, PC1C, PC2C, PD1I, PD1C, PD2C, PNa1I, PNa2C, Pcd1I, Pcd2D, Pht1I, PcdNa1I, PhtD1I, and Pht(Na1I; serving as control sets) generated from the panicles of wild-type rice, all of which were retrieved from Next-Gen Sequence Databases (http://mpss.udel.edu/common/web/library_info.php?SITE=rice_sRNA2&showAll=true) (Nakano et al., 2006). The second group, also obtained from Next-Gen Sequence Databases, contains four data sets (S1dc1I, S1dc1D, S2dc1I, and S2dc1D) generated from whole seedlings of the rice DCL1 RNAi transgenic lines, and 13 data sets (SC1I, SC2I, SC3D, SD1I, SD2D, SNa1I, SNa2D, SCd1I, SCd2D, Sht1I, SCdNa1I, ShtD1I, and Sht(Na1I; serving as control sets) generated from wild-type seedlings. The third group includes GSM520637 prepared from the seedlings of the rice DCL1 RNAi transgenic lines, and three control sets (GSM520638, GSM520639, and GSM520640) generated from the wild-type seedlings, all of which were retrieved from GEO (Gene Expression Omnibus; http://www.ncbi.nlm.nih.gov/geo/) (Barrett et al., 2009).

For each data group, the pre-set criteria were applied to search for the DCL1-dependent sRNAs. As a result, a total of 5898 non-redundant DCL1-dependent sRNAs were identified from the three groups (Fig. 1A; Supplementary Data S1 is available at JXB online). After removing the sRNAs perfectly mapped onto the miRBase-registered (release 20) pre-miRNAs of rice, 5795 sRNAs were retained (Supplementary Data S2 is available at JXB online). To find the genomic loci of these sRNAs, we mapped the 5795 sRNAs onto the whole rice genome, and only the perfectly matched loci were retained. As a result, 2310 DCL1-dependent sRNAs with 3092 genomic loci were identified (Supplementary Data S3 is available at JXB online). In order to identify the hot spots for DCL1-dependent sRNA generation on the rice genome, an ‘sRNA cluster’ was defined: a cluster must be constituted by two or more DCL1-dependent sRNA loci, and should be ≤500 nt. As a result, 4631 clusters were obtained (Supplementary Data S4 is available at JXB online). According to the strand bias of the sRNA loci within a cluster, a specific genomic strand was collected for the following analyses. Interestingly, we observed that, in most cases, a dominant portion of the sRNA loci within a cluster were located on the same strand (see Supplementary Data S4 is available at JXB online).

Since this is a case study, a part of the clusters (669 out of 4631) was randomly selected for secondary structure prediction and manual screening. Secondary structures were predicted by RNAfold using two different algorithms (MFE and centroid) (Hofacker, 2003). As mentioned above, DCL1 treats the precursors with highly complementary long-stem structures as the preferential substrates. In this regard,
we performed a manual screening for the conserved long-stem structures containing DCL1-dependent sRNA loci by comparing the MFE- and the centroid-based structures. Fortunately, 60 clusters with the potential of forming long-stem structures were identified (Fig. 2; Supplementary Table S1, Fig. S1, and Data S5 are available at *JXB* online). Among the 60 clusters, 44 clusters were assigned to 16 different groups. Within the same group, the genomic distributions of the clusters were highly overlapped with each other. In addition, the clusters belonging to the same group possess at least one identical long-stem structure containing DCL1-dependent sRNA loci. The genomic locations of these potential sRNA precursors were investigated using the Genome Browser tool provided by MSU Rice Genome Annotation (release 7; Fig. 2. Examples of the potential long-stem structured precursors encoding Dicer-like 1 (DCL1)-dependent small RNAs (sRNAs) identified in the case study. (A) The potential long-stem structured precursor (covering the genomic region from 9 100 101 to 9,100 599 on rice chromosome 4) encoding four DCL1-dependent sRNAs (marked by red bars). (B) The potential long-stem structured precursor (covering the genomic region from 16 814 831 to 16 815 108 on rice chromosome 4) encoding three DCL1-dependent sRNAs (marked by red bars). (C) The potential long-stem structured precursor (covering the genomic region from 2 409 101 to 2 409 293 on rice chromosome 6) encoding two DCL1-dependent sRNAs (marked by red bars). For (A–C), the EST (expressed sequence tag)- and the RNA-seq (RNA sequencing)-based expression evidence is shown on the right of each panel. The horizontal orange bars denote the genomic regions of the potential sRNA precursors. The secondary structures of the precursors were predicted by RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) (Hofacker, 2003).
http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/ (Yuan et al., 2003). Seven precursors reside within the exons of the annotated rice genes, seven reside within the UTRs (untranslated regions), and a dominant portion (45) reside within the intergenic regions. Only one precursor was identified within the ninth intron of LOC_Os04g28420 (Fig. 2B). In addition, we also observed that certain precursors possess hairpin structures similar to the canonical pre-miRNAs, and both the miRNA and the miRNA* candidates were DCL1 dependent. For example, a cluster sequence covering the genomic region from 16 814 831 to 16 815 108 on chromosome 4 could form a highly complementary long-stem structure. Two DCL1-dependent sRNAs that could form a duplex with 3 nt 3' overhangs were identified on this structure (Fig. 1B). Another cluster covering the genomic region from 2 409 101 to 2 409 293 on chromosome 6 could also form a pre-miRNA-like structure. Also, a complementary duplex with 4 nt 3' overhangs formed by two DCL1-dependent sRNAs were identified on this structure (Fig. 1C). Whether these precursors are genuine pre-miRNAs or some other novel sRNA precursors whose processing requires DCL1 needs further investigation.

As mentioned above, a large portion of the identified sRNA precursors resided within the intergenic regions of the rice genome. It is desirable to inspect the expression levels of these precursors, since the transcription of these genomic regions is the prerequisite of forming long-stem RNA structures. By utilizing the EST and RNA-seq data provided by the Genome Browser tool of MSU Rice Genome Annotation (release 7), expression evidence was obtained for 55 out of the 60 sRNA precursors. For the remaining five precursors, two were expressed weakly, and no expression evidence was obtained for the other three precursors (Supplementary Table S1 and Fig. S2 are available at JXB online).

Taken together, the results of this case study indicate that the structure-based workflow is applicable for the identification of the potential precursors encoding DCL1-dependent sRNAs in plants. However, many more efforts are needed to characterize the novel DCL1-dependent sRNA population thoroughly. For instance, to support our proposed model of DCL1-mediated processing of the long-stem structures for sRNA generation, the independence of RDR (RNA-dependent RNA polymerase) activity needs experimental validation. Double-stranded RNA sequencing could be employed as a powerful approach to validate the predicted long-stem structures (Zheng et al., 2010). In addition, to investigate the biological functions of the DCL1-dependent sRNAs, Argonaute (AGO) enrichment analysis should be performed. Then, target prediction and degradome sequencing data-based target validation could be carried out for the AGO1-enriched sRNAs which possess great potential of performing target cleavages (Fig. 1B).

The lack of genomic and functional information of the non-miRNA, DCL1-dependent sRNAs becomes the main obstacle to study this kind of sRNA species which were once discarded as ‘by-products’. Based on the two observations that 5795 out of 5898 DCL1-dependent rice sRNAs identified in the case study could not find perfect loci on the registered miRNA precursors and the accumulation levels of these sRNAs in the control data sets (such as in the wild-type plants) were >3 RPM, it is unbelievable that all of the sRNAs are just ‘by-products’. In a recent review by Axtell (2013), the non-miRNA, DCL1-dependent sRNAs derived from single-stranded hairpin structures were defined as hairpin RNAs (hpRNAs) whose biogenesis pathway(s) and biological roles remain to be investigated. One of the important uses of the workflow proposed in this study is to uncover the hpRNA-like sRNAs with a genome-wide scale. Thus, we hope that our study could inspire more research efforts to study novel sRNA species in plants.

**Supplementary data**

Supplementary data are available at JXB online.

**Data S1.** List of 20–24 nt DCL1-dependent sRNAs identified in rice.

**Data S2.** List of 20–24 nt DCL1-dependent rice sRNAs after removing those perfectly matched to the rice pre-miRNAs.

**Data S3.** A total of 30 0902 perfectly matched genomic loci of 2310 DCL1-dependent rice sRNAs.

**Data S4.** Detailed information of the 4631 sRNA clusters.

**Data S5.** Detailed information of the 60 sRNA clusters capable of forming long-stem structured precursors.

**Figure S1.** RNAfold-predicted secondary structures of the 60 sRNA precursors. Both MFE and centroid structures are included for each precursor.

**Figure S2.** EST- and RNA-seq-based expression evidence for the identified potential sRNA precursors.

**Table S1.** List of identified potential precursors generating DCL1-dependent sRNAs.

**Acknowledgements**

We would like to thank all the scientists behind the publicly available data sets. This work was supported by the National Natural Sciences Foundation of China [31271380], [3110937], and [31371328], and the Starting Grant funded by Hangzhou Normal University to YM [2011QDL60].

**References**


