Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*

Ryan D. Morin,[1] Gozde Aksay,[2] Elena Dolgosheina,[3] H. Alexander Ebhardt,[4] Vincent Magrini,[5] Elaine R. Mardis,[5] S. Cenk Sahinalp,[6] and Peter J. Unrau[3,7]

[1]*Genome Sciences Centre, BC Cancer Agency, Vancouver V5Z 1L3, Canada;* [2]*University of Washington, Department of Genome Sciences, Seattle, Washington 98195-5065, USA;* [3]*Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby V5A 1S6, Canada;* [4]*Department of Biochemistry, University of Alberta, Edmonton T6G 2H7, Canada;* [5]*Washington University School of Medicine, Genome Sequencing Center, St. Louis, Missouri 63108, USA;* [6]*Department of Computing Science, Simon Fraser University, Burnaby V5A 1S6, Canada*

The diversity of microRNAs and small-interfering RNAs has been extensively explored within angiosperms by focusing on a few key organisms such as *Oryza sativa* and *Arabidopsis thaliana*. A deeper division of the plants is defined by the radiation of the angiosperms and gymnosperms, with the latter comprising the commercially important conifers. The conifers are expected to provide important information regarding the evolution of highly conserved small regulatory RNAs. Deep sequencing provides the means to characterize and quantitatively profile small RNAs in understudied organisms such as these. Pyrosequencing of small RNAs from *O. sativa* revealed, as expected, ~21- and ~24-nt RNAs. The former contained known microRNAs, and the latter largely comprised intergenic-derived sequences likely representing heterochromatin siRNAs. In contrast, sequences from *Pinus contorta* were dominated by 21-nt small RNAs. Using a novel sequence-based clustering algorithm, we identified sequences belonging to 18 highly conserved microRNA families in *P. contorta* as well as numerous clusters of conserved small RNAs of unknown function. Using multiple methods, including expressed sequence folding and machine learning algorithms, we found a further 53 candidate novel microRNA families, 51 appearing specific to the *P. contorta* library. In addition, alignment of small RNA sequences to the *O. sativa* genome revealed six perfectly conserved classes of small RNA that included chloroplast transcripts and specific types of genomic repeats. The conservation of microRNAs and other small RNAs between the conifers and the angiosperms indicates that important RNA silencing processes were highly developed in the earliest spermatophytes. Genomic mapping of all sequences to the *O. sativa* genome can be viewed at http://microrna.bcgsc.ca/cgi-bin/gbrowse/rice_build_3/.

[Supplemental material is available online at www.genome.org.]

In plants, small RNAs play an important role in transcriptional and post-transcriptional gene regulation (Bartel 2004) that include viral defense (Wang and Metzlaff 2005), silencing of transposable elements, and general heterochromatin maintenance (Herr 2005). The small RNAs produced by angiosperms such as *Arabidopsis thaliana* can be broadly classified by the mechanism of their maturation and ultimate function. The microRNAs (miRNAs) are cleaved from a stem–loop precursor molecule (Park et al. 2002) by the endonuclease DCL1 and are ~19–24 nt long (Bartel 2004; Jones-Rhoades and Bartel 2004). The other endogenous small RNAs, collectively termed siRNAs, derive from double-stranded RNA precursors that are processed by homologs of DCL2, DCL3, and DCL4 (Vazquez 2006). The heterochromatin siRNAs are a diverse set of 24-nt-long small RNAs that are processed by DCL3 from double-stranded RNA precursors produced by RDR2 (Xie et al. 2004). These RNAs are involved in heterochromatin formation and maintenance by directing sequence-specific DNA and histone methylation of transposable elements and some larger genomic loci (Pontier et al. 2005). Other 24-nt long siRNAs produced by DCL2 in *A. thaliana* can direct an initial cleavage of target transcripts, which are further cleaved into 21-nt siRNAs by DCL1 (Borsani et al. 2005). Finally, the *trans*-acting siRNAs (tasiRNAs), which are 21 nt long, are matured by a poorly understood mechanism involving DCL4. These tasiRNAs perform post-transcriptional gene silencing much like the miRNAs (Xie et al. 2004).

Identification of functional small RNAs in other plant species has, until recently, been accomplished by searching for homologous sequences in expressed sequence data (Zhang et al. 2006a) and genomic sequences (Bonnet et al. 2004) and has been, with a few exceptions (Williams et al. 2005; Talmor-Neiman et al. 2006), limited to the discovery of the more highly conserved families of miRNAs. Recent evidence suggests that the miRNA repertoire of any plant or animal species comprises a set of conserved ancient miRNAs as well as many recently evolved species-specific miRNAs (Lindow and Krogh 2005; Rajagopalan et al. 2006), which would elude detection by most comparative methods. As they are likely under relaxed selective constraint, the nonconserved miRNAs appear to be rapidly evolving (Rajagopalan et al. 2006; Fahlgren et al. 2007).

For this study, we chose to perform a deep sampling of small RNA sequences from the gymnosperm *P. contorta* accompanied by a lighter sampling of small RNA sequences from a previously studied angiosperm *O. sativa* to facilitate the direct comparison of small RNA populations between these two distantly related species. This choice was made based on a recent RNA silencing

[7]**Corresponding author.**
E-mail punrau@sfu.ca; fax (778) 782-5583.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6897308.

survey we performed across the vascular land plants (E.V. Dolgosheina, R.D. Morin, G. Aksay, S.C. Sahinalp, V. Magrini, E.R. Mardis, J. Mattsson, and P.J. Unrau, unpubl.). This survey indicated that the gymnosperms and in particular the conifers have an unusual RNA silencing signature relative to the angiosperms. Specifically, all conifers tested to date have failed to show appreciable amounts of 24-nt small RNA and instead produce substantial amounts of RNA that is exactly 21-nt long. The reason for this unusual change in RNA expression is not fully understood, but may be related to the presence of a new DCL family in the conifers that is suggested by an analysis of available ESTs from these plants. This difference together with the ~350 million years ago (Mya) divergence between conifers and angiosperms prompted us to perform a detailed comparison of the small RNAs in hope of elucidating evolutionarily important small RNA sequences in the plants.

Without experimental support for the existence of the mature miRNA molecule or direct homology with experimentally confirmed miRNAs, any in silico predictions of miRNAs generally do not qualify as candidates for submission to the microRNA registry miRBase (Griffiths-Jones 2006), thus remaining as predictions until they are ultimately sequenced or their expression confirmed by a hybridization-based method (Ambros et al. 2003). Some recent work in this field has attempted to improve miRNA annotation to better handle the type of data currently produced using high-throughput small RNA sequencing strategies (Rajagopalan et al. 2006; Johnson et al. 2007) accomplished by either massively parallel signature sequencing (Nakano et al. 2006), pyrosequencing (Yao et al. 2007), or Solexa sequencing (Morin et al. 2008). Rather than focusing on searching genomic and expressed sequences for miRNAs, the emerging challenge is to sort through diverse sequences from small RNA cDNA libraries and identify the functional classes including, but not limited to, miRNAs that are likely to be biologically significant.

Sequencing-based small RNA discovery produces hundreds of thousands of small RNA sequences with only a small fraction representing known miRNAs (Gustafson et al. 2005; Lu and Tei 2005; Rajagopalan et al. 2006; Johnson et al. 2007). The remaining sequences reveal complete or fragmented noncoding RNAs (ncRNAs; i.e., rRNA, tRNA, snoRNAs, and snRNAs) or messenger RNAs (mRNAs) in addition to diverse populations of siRNAs. With an annotated genome, the former can be readily identified based on their perfect alignment to genomic regions annotated as ncRNAs. Though little is known about the specific siRNAs of plants other than those of *A. thaliana*, conserved small RNAs that are observed in evolutionarily distant plant species will likely prove to be either siRNAs, miRNAs, or fragments of larger ncRNAs involved in important aspects of cellular regulation.

Results

Small RNA sequencing and sequence processing

A total of 142,493 and 11,436 sequence reads were obtained from the *P. contorta* and *O. sativa* libraries, respectively. After removing artifacts (see Methods), 130,998 (92%) and 11,329 (99%) sequences remained for analysis. The sequence artifacts include products of multiple adapter ligation or "empty" constructs, in which the two adaptors ligated to one another without containing a small RNA. There were a total of 58,466 (44.6% of data set) and 8615 (76.0% of data set) unique sequences in the *P. contorta* and *O. sativa* libraries, respectively. Of these, 11,375 *P. contorta* and 707 *O. sativa* sequences were counted at least twice in the libraries, leaving 47,091 and 7908 singletons (35.9% and 69.8% of the total sequences, respectively). This observation suggests that a huge diversity of small RNA sequences existed in each library.

For *P. contorta*, the lengths of small RNAs ranged from <11 nt to >30 nt, and the distribution of unique small RNA length is summarized in Figure 1 (black bars). The *O. sativa* sequences spanned the same range of lengths; however, the overall distribution of these lengths was strikingly different (Fig. 1, light gray bars), with one major peak at 24 nt and another minor peak at 21 nt. Of all the length classes, the 24-nt fraction was the most diverse in *O. sativa*, with only 305 (16.5%) of the 24-nt sequences sharing high sequence similarity with at least one other sequence

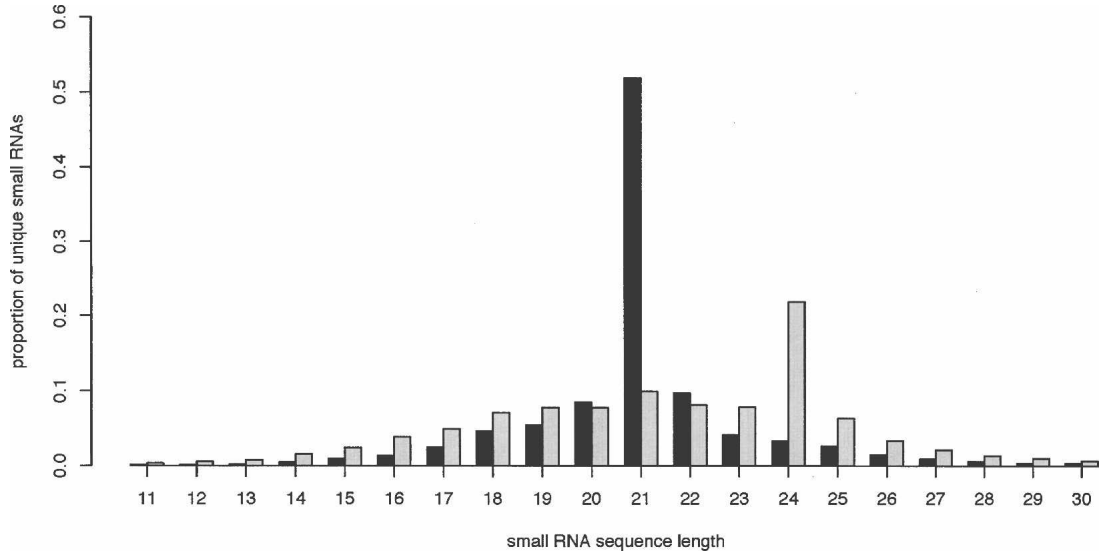

Figure 1. Lengths of unique small RNA sequences from *P. contorta* (black bars, 58,466 sequences) and *O. sativa* (gray bars, 8615 sequences). The bulk of *P. contorta* small RNAs are 21 nt long with low variance (σ = 8.1). The rice sequences have a major peak at 24 nt and a minor peak at 21 nt, yielding a median of 22 nt and a higher variance (σ = 20.0). The 21-nt peak becomes more prominent when sequence degeneracy is considered (not shown).

in that population based on a clustering analysis (see Methods). The relative representation of 24-nt RNA in *P. contorta* was small (2.5%) in comparison to *O. sativa* and had a lower diversity, with 36% of the 24-nt sequences sharing high similarity with another sequence. In contrast, the 21-nt RNAs obtained from *P. contorta* were more diverse than the *O. sativa* 21-nt sequences, comprising a total of 29,924 unique sequences. This suggests an expansion of miRNA families in the gymnosperms, diversification of other 21-nt RNA producing pathways, or a functional replacement of most of the heterochromatin siRNAs by 21-nt sequences. In either case, this observation reveals a significant difference in the small RNA biogenesis pathways of the angiosperms and gymnosperms. The apparent absence of 24-nt small RNA in the *P. contorta* sequencing data is entirely consistent with our survey of the vascular plants, which found no evidence for 24-nt RNA expression in the conifers, as judged by the direct 5′ end labeling of total RNA extracts (E. Dolgosheina, R.D. Morin, G. Aksay, S.C. Sahinalp, V. Magrini, E.R. Mardis, J. Mattsson, and P.J. Unrau, unpubl.).

Genome mapping and small RNA annotation—*O. sativa* small RNAs

Of 8615 unique *O. sativa* small RNA sequences, 3814 had at least one perfect alignment in the *O. sativa* genome. Small RNA sequences were annotated as one of six broad groups based on their overlap with *O. sativa* genome annotations (Itoh et al. 2007) or their alignment to sequences in Rfam (Griffiths-Jones et al. 2003) (miRNA, repeat-derived siRNAs, tRNA, rRNA, snRNA/snoRNA, or un-annotated). The smallest group, which was excluded from further analysis, contained the small nuclear and small nucleolar RNAs (snRNAs and snoRNAs). This group comprised 63 sequences from *O. sativa* and 646 from *P. contorta*. While the majority of annotated sequences mapped to the *O. sativa* genome only once, subsets of small RNAs when separated by type mapped to the genome with interesting and distinct distributions (Fig. 2; Supplemental Fig. 1). Our classification supports the notion that the 21-nt fraction of the *O. sativa* small RNA sequences (Fig. 2B) includes members of 18 conserved miRNA families (summarized in Supplemental Table 1), whereas the 24-nt fraction was dominated by small RNAs derived from genomic repeats and intergenic regions (Fig. 2A,E), suggesting they are mainly acting as heterochromatin siRNAs. Many of the sequences classified here as miRNAs could not be unambiguously assigned to one miRNA gene, as members of many miRNA families share a common, or highly similar, mature miRNA sequence. For example, the sequence TGAAGCTGCCAGCATGATC could belong to any of nine current members of the MIR167 family. Further, there was not a direct one-to-one relationship between RNA sequences and miRNAs, with most of the miRNA genes apparently producing multiple variants (termed isomiRs here) that resulted from slight variability in the DCL cleavage sites or subsequent processing/degradation leading to removal of terminal nucleotides (example in Supplemental Fig. 2b).

As the majority of *O. sativa* sequences were 24 nt in length, many of which were derived from genomic repeats, we assumed that the library contained a large proportion of siRNAs generated by a DCL3-mediated pathway similar to that in *A. thaliana* (Xie et al. 2004). In an attempt to identify genomic target sites of the *O. sativa* siRNAs, all small RNA sequences were aligned to the *O. sativa* genome allowing some degeneracy (Methods). This approach should highlight all sites in the genome to which a given small RNA could readily anneal; 6859 (80%) of the *O. sativa* sequences aligned at least once by this method. This suggested that many of the unmapped sequences were not the result of simple sequencing errors. Since this degenerate alignment method allowed for up to three mismatches, only reads with multiple errors or belonging to unsequenced regions of the genome should not be mapped. A histogram of small RNA alignment frequency for 21- and 24-nt small RNAs along the length of each of the 12 nuclear chromosomes is shown in Figure 3, demonstrating that the small RNA alignments were nonuniform. It is known that heterochromatin siRNAs can elicit DNA methylation and histone modification at partially complementary regions of chromosomal DNA at repetitive elements (Herr 2005) as well as rRNA loci (Xie et al. 2004). The alignment distribution allows a global visualization of the regions potentially targeted for modification by siRNAs. These sites include many of the rRNA clusters and most of the known centromeres, the latter of which are known to contain a specific repertoire of genomic repeats (Cheng et al. 2002).

Partitioning small RNAs into clusters of similar sequences

Annotation of small RNAs from *P. contorta* was precluded by the lack of any genomic sequence. Further, the direct comparison of these sequences to known miRNAs allows the identification of only the highly conserved miRNA family members. Clustering of *P. contorta* and *O. sativa* small RNA sequences based on sequence similarity was performed to enable identification of related, but not identical, sequences. Also, we assume the observed low success rate in mapping *O. sativa* sequences to the parent *O. sativa* genome likely results from a combination of systematic and sporadic sequencing errors, polymorphisms, RNA editing, and errors in the reference genome. Clustering highly similar sequences provides a mechanism to resolve this issue. The most frequently observed sequences in a cluster are likely to represent real sequences, and the more rare variants likely represent sequencing errors or reads deriving from polymorphic regions. Any sequence that was not mapped to the *O. sativa* genome due to evolutionary divergence (for *P. contorta* sequences) or the aforementioned reasons (*O. sativa* sequences) could still be annotated based on its presence in a cluster with an annotated sequence. We performed sequence-based clustering of all *O. sativa* and *P. contorta* sequences along with those known miRNA sequences from miRBase (see Methods). The result was 4722 clusters, ranging in size from two sequences (2366 clusters) to 4386 sequences (one cluster). A total of 20,434 *P. contorta* (of 58,466 unique) and 3959 *O. sativa* (of 8615 unique) sequences reside in these clusters; 4511 of the clusters contained at least one *P. contorta* sequence, whereas only 547 contained at least one *O. sativa* sequence, with 373 clusters comprising sequences from both species. These clusters were therefore very likely to represent conserved classes of either small RNAs or recurrent degradation fragments of larger ncRNAs.

The quality of each cluster of sequences was assessed by calculating the mean information content of the multiple sequence alignment of its sequences (Schneider and Stephens 1990). The information content is a function of the entropy of each position in the alignment, with more insertions/deletions (indels) and discordant sites decreasing the mean information content of an aligned cluster. The mean information content of a cluster was roughly dependent on the number of sequences it contained (Supplemental Fig. 3), with some variability caused by incomplete overlap of aligned sequences and the number of insertions and deletions between sequences within a cluster. With

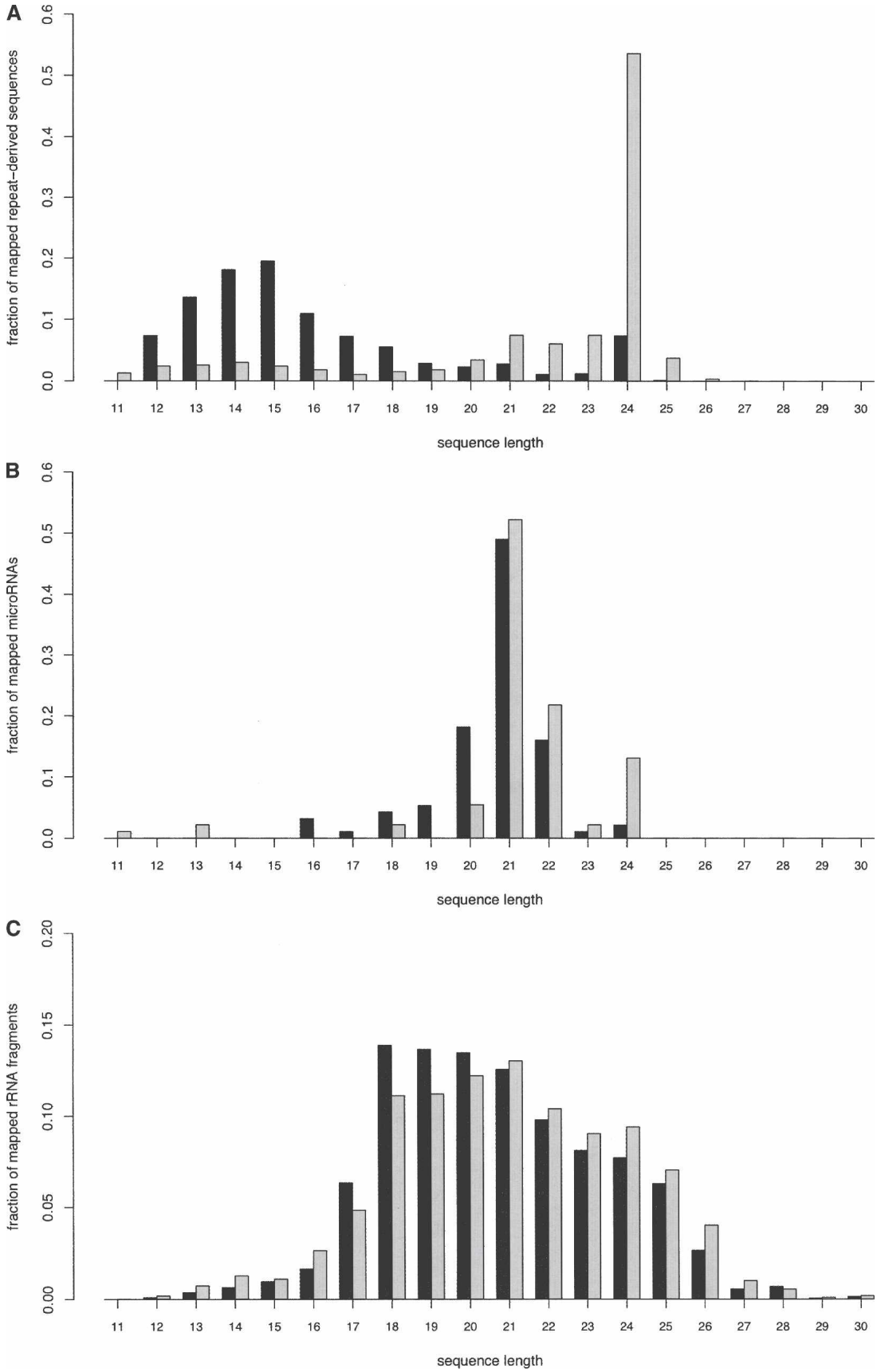

Figure 2. (Continued on next page)

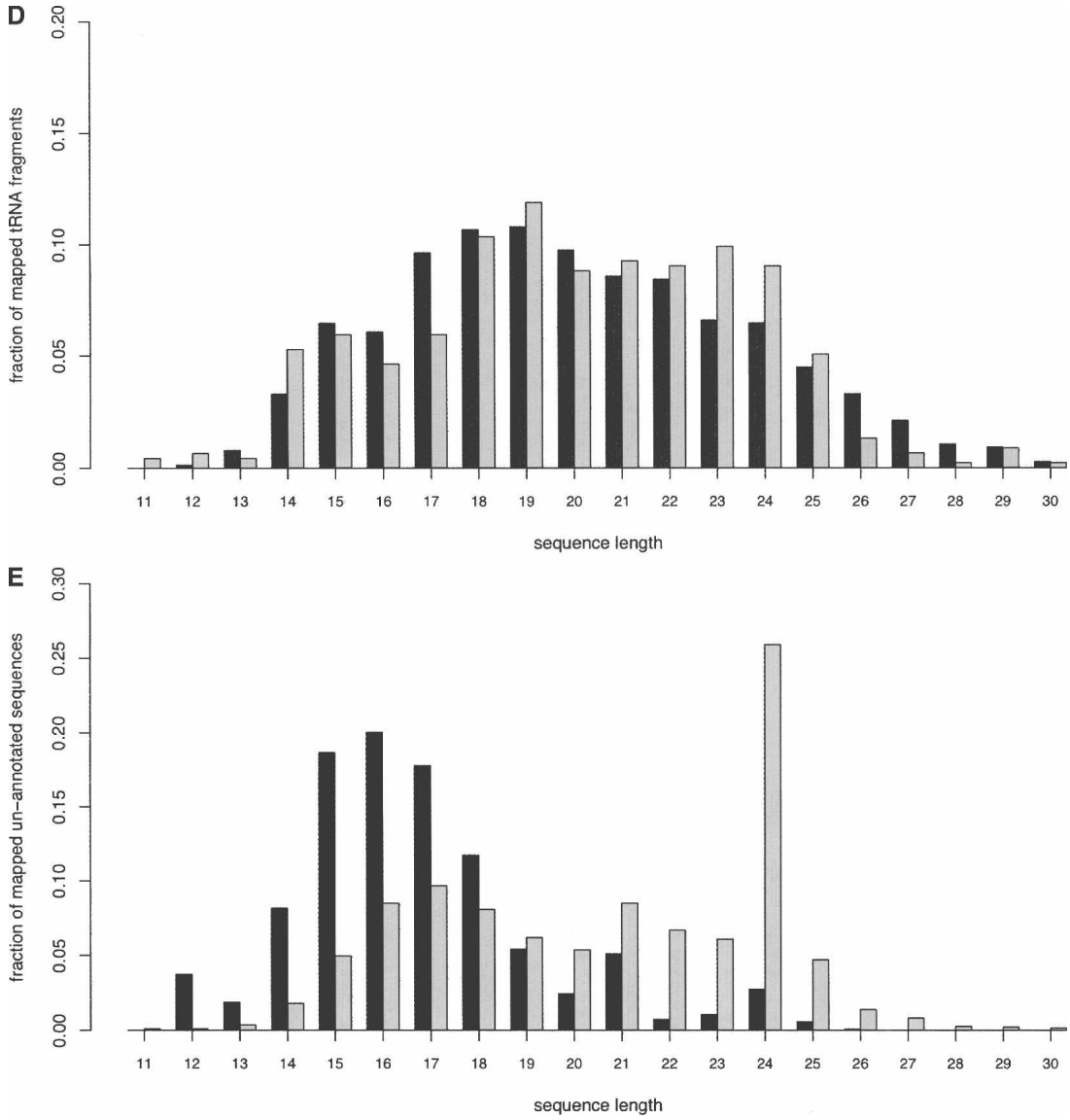

Figure 2. Length distribution of unique *P. contorta* (black bars) and *O. sativa* (gray bars) small RNAs sorted by class and that map perfectly to at least one genomic locus in the *O. sativa* genome. A total of 5129 unique *P. contorta* sequences mapped to the *O. sativa* genome. The *Y*-axis reflects the fraction of sequences of the stated annotation residing within each length bin. (*A*) Genomic repeats as annotated by RepeatMasker using RepBase, v. 9.04. Two distinct lobes are observable, the first due mainly to the high probability of shorter sequences occurring by chance in the genome. The second lobe peaks for both species with 24-nt sequences, suggesting an over-representation of conserved sequences that are 24 nt in length. (*B*) Conserved miRNAs matching sequences in miRBase release 9.1. These miRNAs are almost exclusively 20–22 nt in length in both species. (*C*) rRNA as identified by overlap with annotated rRNA genes in the rice genome. The relative lack of ribosomal fragments below 18 nt likely corresponds to the cutoff imposed by gel purification. (*D*) tRNA as annotated by tRNAscan-SE. (*E*) Un-annotated, small RNAs not overlapping with sequences annotated by the aforementioned methods. Again, a peak is observed at 24 nt for *P. contorta* sequences, revealing a higher proportion of evolutionarily conserved sequences in this class reside in the 24-nt fraction than in the 23- or 25-nt fractions.

this clustering method, sequences of known miRNAs generally clustered within their miRBase family. Supplemental Table 1 summarizes the conserved miRNA families represented by these clusters. In some cases, multiple families resided in the same cluster due to similarity between one or more sequences within the families. For example, one cluster contained sequences from MIR165 and MIR166, while another contained sequences from MIR319 and MIR159. This was not surprising as alignment of sequences between these families results in three or fewer mismatches. Homologous miRNA genes can result in the production of miRNAs that differ only at the 5′ or 3′ termini. As mentioned, many of the miRNAs have multiple isomiRs, reflecting variability in miRNA maturation or later processing steps. Our sequence-based clusters containing known miRNAs demonstrated overall higher information content (Supplemental Fig. 3b,c) than the clusters of the same size but comprising degradation products of larger ncRNAs. This is a reflection of the high sequence conservation of miRNAs within the same family as well as the presence of isomiRs with similar yet distinct sequences (Morin et al. 2008). These observations were the basis for our application of a machine learning method to predict candidate miRNA families from un-annotated clusters (discussed below).

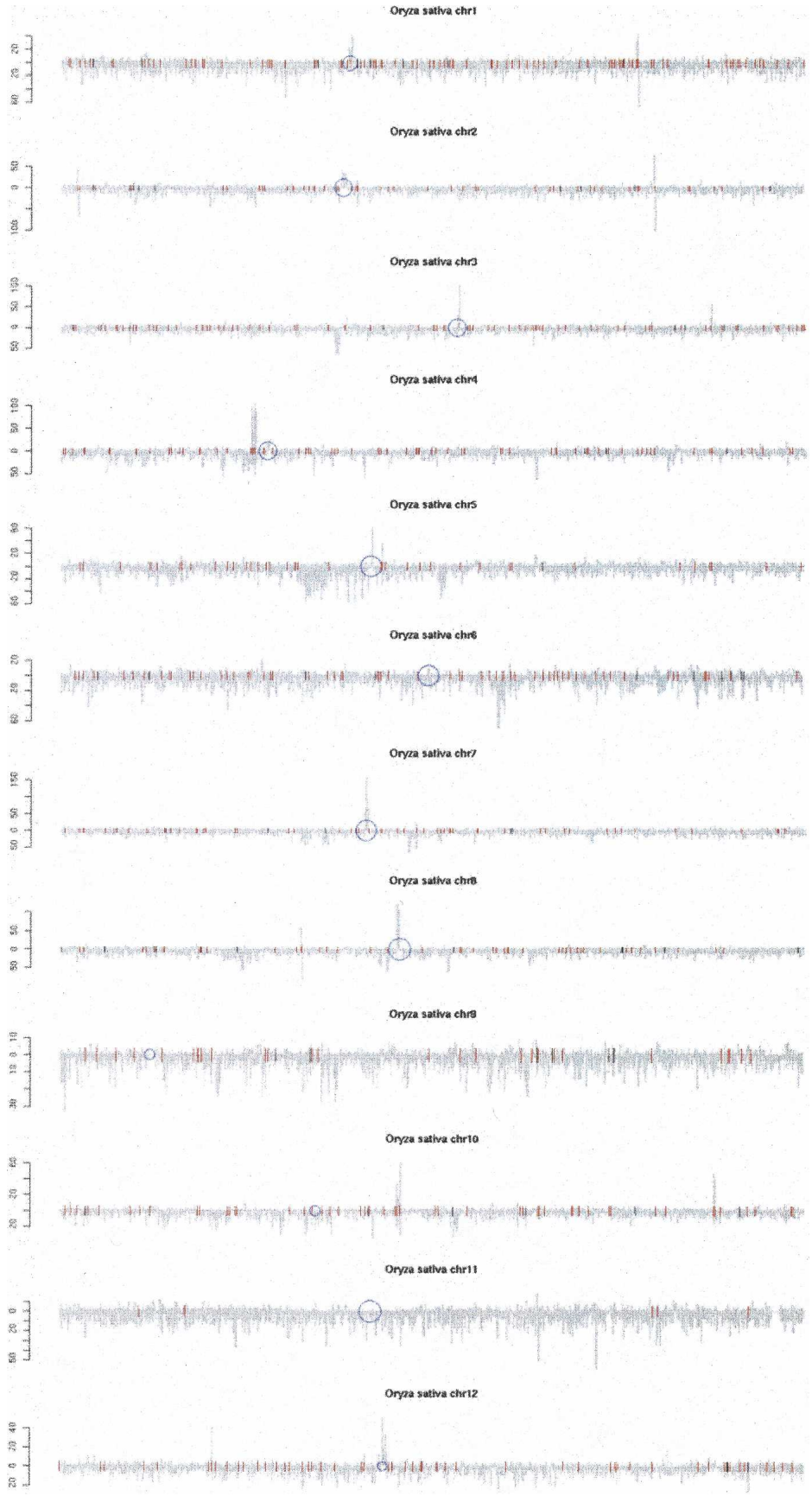

Figure 3. (Legend on next page)

Identifying miRNA candidates specific to *P. contorta* or *O. sativa*

A specific pre-miRNA secondary structure is necessary to establish a small RNA as a miRNA candidate. Owing to the lack of any *Pinus* genomic sequence, *P. contorta* small RNA sequences were aligned to EST sequences originating from *Pinus taeda*, a closely related species of pine, while *O. sativa* sequences were aligned to the publicly available *O. sativa* EST sequences. The number of UniGene clusters (Pontius et al. 2002) (~14,000) can be used as a rough estimate of the number of *P. taeda* genes currently represented by expressed sequences in GenBank. Hence, these cDNA and EST sequences (collectively referred to as ESTs) are not a full representation of the transcriptome as compared to some other EST libraries for gymnosperms (Quackenbush et al. 2001; Pavy et al. 2005). A total of 13,396 (10.2%) *P. contorta* small RNA sequences had perfect alignments to one or more *P. taeda* EST; 318 of the sequences had multiple semi-overlapping alignment regions with numerous distinct small RNAs, a signature that would suggest these derived from degradation of larger ncRNAs (see Supplemental Fig. 2a).

Table 1. Small RNA clusters annotated as miRNA families by EST alignments and folding properties

miRNA family	Representative sequence	Total parent ESTs	Count (most abundant in cluster)	Species
MIR950	**TCACGTCAGGTCCTCGATGGTT**[a]	1	262	*P. contorta*
MIR1309	TGATGGCTTTTCTGAAGGACA[b]	3	11	*P. contorta*
MIR946	CAGCCCTTCTCCTATCCACAAC	2	152	Both
MIR1310	GGCATCGGGGGCGTAACGCCCCT	4	20	Both
MIR1311	TCAGAGTTTTGCCAGTTCCGCC	1	4	*P. contorta*
MIR1312	TTTGGAGAGAAAATGGCGACAT[a]	2	43	*P. contorta*
MIR1313	TACCACTGAAATTATTGTTCG[a]	1	15	*P. contorta*
MIR1314	TCGGCCTTGAATGTTAGGAGAA	3	125	*P. contorta*
MIR1315	TGGAGGCCTGTCAGGTTCCCA[a]	4	34	*P. contorta*
MIR946	TGTGGATAGAGAAGGGTTAGT	2	39	*P. contorta*
MIR1316	TCCATACACAAACCATTGGAA	1	34	*P. contorta*
MIR1317	TAGGGAACCCCATTCCCATAAA	1	5	*O. sativa*
MIR1318	TCAGGAGAGATGACACCGAC	1	2	*O. sativa*
MIR1319	AACCGGCATCTGTAATATATTATA	1	1	*O. sativa*
MIR1320	TGGAACGGAGGAATTTTATAG	1	1	*O. sativa*
MIR444	GGCTTTCTTGCAAGTTGTGCA	1	1	*O. sativa*

Bold sequence indicates that expression was confirmed by Northern analysis.
[a]Likely miR* sequences for this miRNA were also observed.
[b]Also identified by the SVM method (present in Table 2).

Some *P. taeda* and *O. sativa* small RNAs could be aligned to their parent EST sequences in only a few discrete regions, a characteristic shared with many of the known miRNA sequences in these libraries (see Supplemental Fig. 2b). Three hundred thirty-six EST sequences had small RNA alignment patterns that were deemed similar to pre-miRNAs and were carried forward for folding analysis (Methods). Along with candidate novel miRNA genes, this procedure correctly identified three sequence clusters belonging to known miRNA-families (MIR166, MIR396, and MIR159). The novel miRNA sequences as well as the EST sequence folding results are summarized in Table 1 and in the Supplemental tables. Two examples of these novel miRNAs are also included in Figure 4, and the supplements include the folded structure of all candidate pre-miRNAs and a sequence logo representing a consensus of each cluster of novel miRNA sequences. Potential homologs, noted in Figure 4, were identified by searching miRBase (v. 10.0) using the SSearch tool.

Not surprisingly, the putative miRNA families (clusters) identified in this process typically displayed above-average information content (Supplemental Fig. 3), following the trend of the known miRNAs previously identified. Two of the clusters identified as comprising miRNAs by alignment to *P. taeda* sequences also included small RNA sequences obtained from *O. sativa*, suggesting the identification of two novel conserved miRNA families by this process (Table 1). From the alignment of our small RNAs to *O. sativa* EST sequences, a few novel *O. sativa* miRNAs not found in the *P. contorta* library were also identified. Because this analysis was based on the data from the previous release of miRBase, these novel miRNAs can be compared to a more recent set of known miRNAs. Strikingly, a search of the current miRBase (v. 10.0) sequences reveals potential homologs for all but two of these sequences. Based on these results, this method appears to provide a fast and reliable way to identify putative novel miRNAs from this type of data using the a priori assumption of variable mature sequence length (isomiRs) and degradation of the remaining structural fragments of pre-miRNAs (Ambros et al. 2003).

Many clusters of *P. contorta* specific sequences could not be aligned to any *P. taeda* EST. This can be partially attributed to the limiting number of *P. taeda* ESTs currently available in GenBank. Identification of nonconserved *P. contorta* miRNAs in those remaining clusters relied on other characteristics besides the standard method of annotation (i.e., sequence folding and hairpin characterization). Our observation that miRNAs cluster to one another in a distinctive way suggested that the characteristics of a cluster—for example, its size, length, and information content—would allow classification of unknown clusters as miRNAs, bypassing the necessity of full pre-miRNA sequences for annotation. A statistical machine learning approach known as a support vector machine (SVM) was used to test this hypothesis (see Methods).

When our SVM classifier was applied to all sequence clusters comprising sequences classified as "un-annotated," 44 were predicted to represent novel miRNAs (Table 2); 19 of these clusters contained at least one sequence with an alignment to a *P. contorta* EST sequence, allowing validation using the aforementioned method of RNA folding and structure evaluation. Of these, five aligned perfectly to one or more EST that could fold into a signature hairpin structure with the nucleotides yielding the small RNA positioned within the stem region (Fig. 4; Table 2; Supplemental Tables), thus confirming they were miRNAs by current standards. One of these miRNAs was also found by the previous annotation method (based on alignment of small RNA sequences to ESTs), but the remaining four were missed by that

Figure 3. Degenerate alignment density of *O. sativa* 24- and 21-nt small RNAs to the nuclear genome. All distinct *O. sativa* small RNA sequences were aligned to all 12 chromosomes allowing degenerate alignments (see Methods). The 24-nt (positive axis) and 21-nt sequences (negative axis) demonstrated distinctive alignment patterns. Specifically, "hot spots" of 24-nt sequences include the heterochromatic regions including most centromeres (blue) as well as some clusters of ribosomal RNA genes (red). Positions of known miRNA genes are also marked (black).

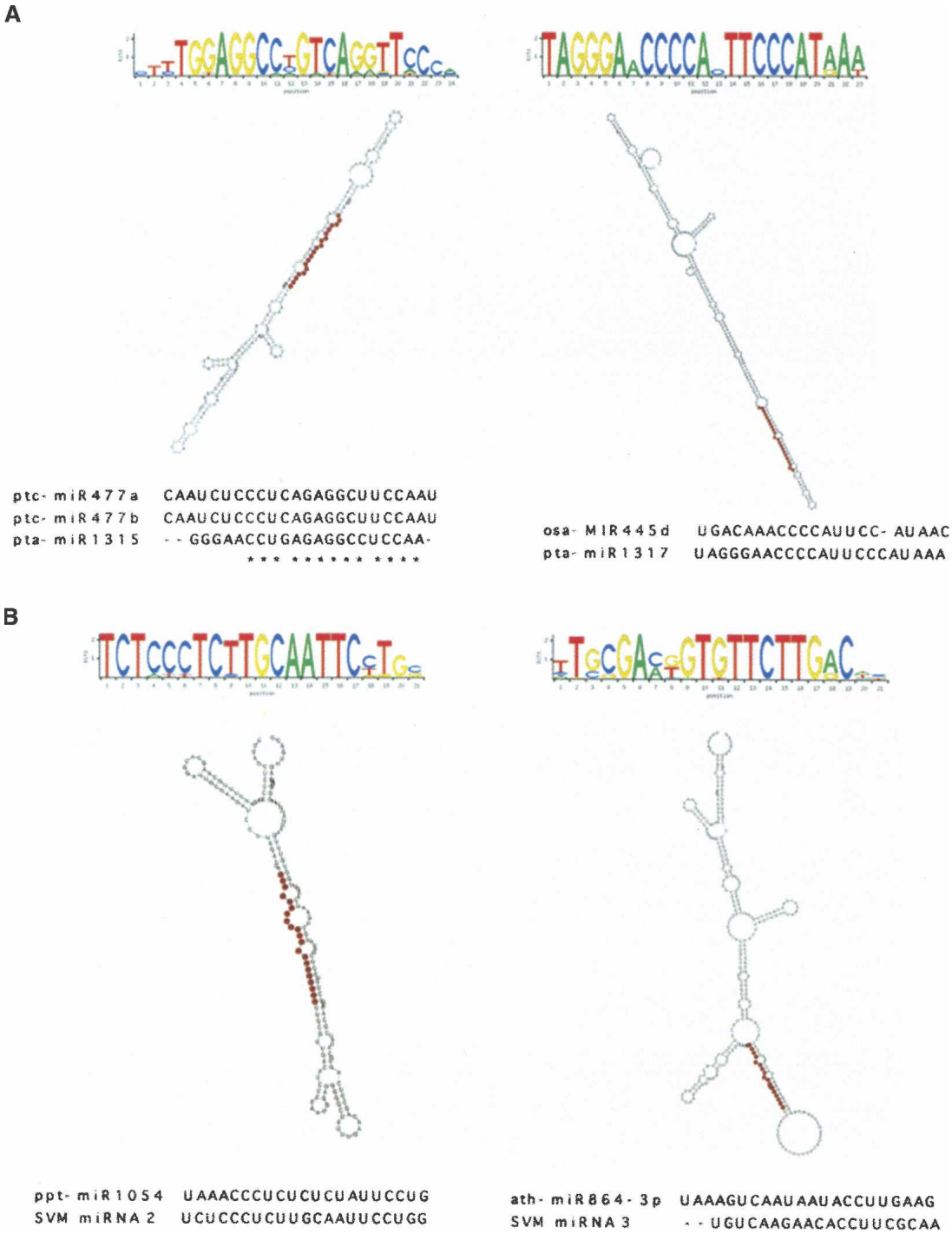

Figure 4. Summary of four novel miRNA clusters identified by two separate methods. Structures and sequence logos are shown for four novel miRNA clusters identified by either EST folding (*A*) or support vector machine (SVM) techniques (*B*). Sequence logos were generated from the multiple-sequence alignment of all small RNA sequences in the cluster. Folded structure (as predicted by RNAFold) includes positioning of the mature miRNA sequence (highlighted in red). The most abundant (highest count) sequence from each cluster was considered the most reliable representative mature miRNA and was used to search for EST (for those identified by SVM) and potential homologs in miRBase. The alignment of each of these sequences to either its best hit or hit(s) in miRBase is included.

approach because they only had a single representative small RNA that aligned to that EST; hence, no isomiRs were sequenced for this miRNA. To validate the remaining putative miRNA sequences, one would require additional expressed sequence data (facilitating more folding analyses) or genomic sequences.

Experimental verification of putative miRNA expression

While it is difficult to functionally verify a new miRNA, we sought to confirm the presence of a limited set of some of the small RNAs presented here using a hybridization-based method.

Table 2. Small RNA clusters annotated as miRNA families by SVM classifier

Most abundant sequence in cluster	Best candidate homolog	Total ESTs	Count for most abundant sequence	EST with good structure
TAGAAGATGGCGATATGGACA	osa-miR811a	4	24	CO159415
TCTCCCTCTTGCAATTCCTGG	ppt-miR1054	1	40	CD024998
TTGCGAAGGTGTTCTTGACA	ath-miR864–3p	1	18	DR054135
GTCAAGAACATTTTACCCTTC[a]	ptc-miR399e	2	20	DR692119
TGATGGCTTTTCTGAAGGACA[b]	—	1	11	CX648707
TGAGATGTAACCAAAATTAAG	—	4	8	—
AGCTCTTGGGTGCATTTTCTC	—	0	26	—
GTCGTATTATCAACTATTTCCCAA	—	0	448	—
TCGCTAGAATCTTTGGTCATC	—	0	38	—
TGAGAACAAGTTCATCCAGGC	—	1	56	—
CATCTCTACCGTCTCCTGGTA	—	3	83	—
TCTGCGTTGATAGCTCATCAC	—	0	111	—
ATCGGATTTGATGGCCTTTTTTGAA	ppt-miR1037	0	11	—
TGAAATCTCCAAACTAAGCTC	ath-miR853	0	42	—
TGGTCGTGCATTCTTATACAG	smo-MIR1103	1	22	—
GAAGGCTCAGCTCGATAAGGC	—	1	114	—
GATCTCGCCATCCATGTTGTA	—	3	18	—
CTGGGAACTCTAAATATGGACA	ppt-miR904b	0	20	—
ATCATTCCCATCTAAGATTGG	—	0	3	—
ACTCATTCAATGGAGAGGCAT	—	0	5	—
TGGGTTGCTCTGTTGAAGGCT	ath-miR857	0	12	—
TCCGTCCCATATTTTGATGGC	osa-miR818d	1	29	—
TCAATAGGACATGCCAGCGAA	—	1	32	—
GTGGGAAAGAATTTCATGGTC	—	0	18	—
TCAAGGCCGATATTCTGTCAT	ptc-miR156h	0	18	—
TGAGTTCTTGATTCGCTATAGT	—	0	72	—
AGGGTAAAATGTTCTTGACA	—	3	11	—
TGGACGGATATTCTAGTTGTA	ptc-miR478a	0	14	—
ACTTGAAAATGAGTTCTGGC	—	1	7	—
TCCTCTAAAAATACGGACAAC	ptc-miR478i	0	92	—
AGAGATCTTGATCTCGAGCTC	—	0	29	—
GAAGAGCCTACGACGAGAAGAGCCTACGACA	—	0	2	—
CTGGTCCTTTGAAGCTCAGAC	—	1	6	—
TCGTCTAAAACTATTAGAGCG	—	1	24	—
CGGAGTCTGTGAGCGATTCGA	—	0	11	—
TCAAGCGCTGATAGAGCTCTT	—	0	24	—
TGGTTGTAACAAGGTATCAGC	—	0	18	—
GGAGTAGAGGTTGGCTAGGGC	—	1	17	—
TGCGATTTGCCTTGCCACAAT	—	1	33	—
GGGAGCTGTCTCTAACAGCGC		0	44	—
TCGACAAGCACGTGAGGCAGC	—	0	8	—
TGAGCGCAGCACAGGATGGAA	—	0	6	—
ACTCTTTTCTTTTATAGCAAC	—	0	7	—
TCTTCTCAAATAGATCAGGCT	ath-miR830*	0	13	—

Bold sequence was verified by Northern blot analysis.
[a]Likely miR* sequences for this miRNA were also observed.
[b]Also identified by EST alignment and folding method (present in Table 1).

We queried three RNA sequences (as indicated in Tables 1 and 2 and in Supplemental Table 1): miR396, which was found to be perfectly conserved between *O. sativa* and *P. contorta*, together with two sequences classified as novel miRNAs. All of these sequences were present in the small RNA fraction of *P. contorta* as judged by Northern analysis (Supplemental Fig. 6). This simple screen indicates that RNAs with sequences identical to, or highly similar to, those found in our sequencing analysis must exist in *P. contorta* and lends support to our mode of analysis.

Classifying other *P. contorta*/*O. sativa* conserved small RNAs

Though focus was placed on miRNAs in this study, the bulk of the sequences produced in this study, 57,619 from *P. contorta* and 8486 from *O. sativa*, cannot be shown to represent conserved or novel miRNAs by any of our annotation methods. Conservation of small RNA sequences between a gymnosperm and an angiosperm provides strong support that the sequences regulate highly conserved RNA-mediated processes. The *P. contorta* sequences that could be aligned perfectly to the *O. sativa* genome or to *O. sativa* small RNAs are of particular interest as they represent highly conserved small RNAs in two distantly related organisms. Using the perfect-match alignment approach described for the *O. sativa* sequences, all *P. contorta* small RNA sequences were aligned to the *O. sativa* genomic sequence. A total of 3567 (2.5%) *P. contorta* sequences >16 nt in length aligned perfectly to the *O. sativa* genome (5129 sequences if shorter sequences are included). Of these sequences, only 91 corresponded to known miRNAs, while 632 were deemed fragments of tRNAs, and 2117 fragments of rRNAs.

The remaining 727 *P. contorta* sequences with perfect matches in the *O. sativa* genome correspond to loci of unknown function, though 253 of them correspond to annotated genomic repeats, implicating them as possible conserved repeat-derived heterochromatin siRNAs. The apparent lack of 24-nt sequences

in the *P. contorta* small RNA sequences provoked us to ask whether these heterochromatin siRNAs belong to a less diverse class of *P. contorta* 24-nt sequences masked by the overwhelming dominance of the 21-nt RNAs expressed in this species. The histogram of the lengths of the *P. contorta* small RNA sequences that mapped to *O. sativa* repetitive elements highlights that intergenic and repeat-derived sequences contained a higher fraction of perfectly conserved 24-nt RNAs than those either 23 or 25 nt in length (Fig. 2A,E). These *P. contorta* sequences with perfect matches in the *O. sativa* genome comprise at least 66 distinct repeat-derived siRNAs. For some of these, corresponding small RNAs were sequenced from *O. sativa* as well (see below). In contrast, the majority of the perfectly conserved *P. contorta* 24-nt sequences appeared to derive from both the sense and antisense strands of rRNA genes, supporting the notion that rRNA can be processed into functional siRNAs (see example in Supplemental Fig. 2a) and that this mechanism of small RNA-mediated regulation is conserved in the gymnosperms.

Those sequence-based clusters comprising small RNAs from both species facilitated further identification of conserved *P. contorta*/*O. sativa* 24-nt RNAs while allowing for some sequence divergence. Twenty-six of these clusters had median lengths of ~24 nt and contained sequences from both species. These clusters (Table 3) were mostly rRNA sequences, but a few were tRNA or repeat-derived. The *P. contorta* sequences within these clusters showed a slightly higher variability in sequence lengths (*O. sativa* mean length variance = 4.45, *P. contorta* mean length variance = 5.15). The median length of sequences amongst these clusters was slightly lower for the *P. contorta* sequences (22 nt) as compared to those from *O. sativa* (24.3 nt). This result suggests that these may represent homologs of *O. sativa* heterochromatin siRNAs, many of ribosomal RNA origin, that deviate from a strict 24-nt length criterion.

The list of 727 *P. contorta* small RNA sequences perfectly conserved in *O. sativa* was semi-redundant. Many of these sequences overlapped the same genomic loci either partially or completely. By grouping sequences that align to at least one shared genomic site, this set could be further subdivided into six unclassified distinct groups of small RNAs. Each of these groups was of interest because all the sequences in them mapped to a common set of genomic loci and did not seem to derive from any known ncRNAs. In the first group, the small RNAs consistently aligned to a discrete position in a subset of the LTR-type transposable elements in the *O. sativa* genome. None of these perfectly conserved small RNAs appeared to be miRNAs by current classification standards; however, the consistent observation of sequences from a few loci within these structures also makes them distinct from the usual phased pattern of siRNAs. Intriguingly, the region of the SZ-37/Osr10 LTR repeat (McCarthy et al. 2002) that appears to encode this small RNA has a near-perfect match to many of the other known *O. sativa* repetitive elements (Supplemental Fig. 4). This suggests the possibility that this site may be a potential target of this siRNA in many of the *O. sativa* LTR elements, thus evolutionarily constraining this short region of LTR elements.

In two separate groups of conserved small RNAs, the sequences always appeared to derive from a region upstream of a nuclear-inserted copy of a chloroplast tRNA gene, suggesting a role in, or byproduct of, tRNA maturation. A similar observation has recently been made in *Arabidopsis*, and it was suggested that these sequences are the sequence leaders cleaved during tRNA maturation (Rajagopalan et al. 2006). Notably, however, most of the small RNAs of this type appear to derive from the opposite strand of the tRNA transcript, as do many of the small RNAs that overlap with the neighboring tRNA annotation (Supplemental Fig. 5a). The three remaining groups of small RNAs also appear to derive from chloroplast genes or their nuclear-encoded counterparts. In one case, all small RNAs aligned to one discrete region of the dicistronic chloroplast transcript of the *ndhB* and *rps7* genes. The position of these small RNAs was situated over the site of enzymatic cleavage of this transcript (Hashimoto et al. 2003) just upstream of the *ndhB* start codon (Supplemental Fig. 5b). This site is perfectly conserved in *A. thaliana* and *Zea mays* as well, suggesting its importance and a high likelihood of small RNA

Table 3. Small RNA sequence clusters comprising ~24-nt RNAs from *O. sativa* and *P. contorta*

Cluster identifier	Mean information content of cluster	No. of *O. sativa* sequences	No. of *P. contorta* sequences	Mean *O. sativa* sequence length	*O. sativa* sequence length variance	Mean *P. contorta* sequence length	*P. contorta* sequence length variance	Annotation of cluster
296759	1.34	1	4	24	0	21	0	Repeat
287972	0.88	1	20	24	0	21.2	0.72	Repeat
286067	0.83	2	15	24	0	22	8.14	Repeat
291886	1.81	4	3	24	6.66	20.6	0.33	rRNA
285484	1.42	2	6	24	0	21.5	5.5	rRNA
291079	1.68	2	8	24	0	21.6	1.41	rRNA
283734	1.29	3	14	24	0	22.5	4.88	rRNA
283021	1.48	4	31	24	2.66	22.7	4.98	rRNA
283290	1.11	3	59	24	7	24	6.55	rRNA
284515	1.14	2	25	24.5	4.5	24.1	7.44	rRNA
293185	1.34	4	6	24.7	6.25	23.8	7.76	rRNA
287854	1.67	2	2	25	2	20	0	rRNA
282962	1.27	3	15	25	21	22.2	2.6	rRNA
282880	0.96	6	61	25.6	2.66	23.1	5.42	rRNA
283517	0.57	3	96	27	1	22.9	7.29	rRNA
283489	0.8	7	85	28.2	2.23	24.7	9.38	rRNA
283140	2.01	3	2	24	0	22	8	tRNA
289230	1.42	4	26	24.2	5.58	23.3	5.75	tRNA
282956	0.8	43	67	24.4	9.68	24.4	10.8	tRNA
283258	0.95	17	83	25.1	6.86	23.6	5.15	tRNA
283074	1.54	11	33	25.8	6.16	24.6	7.8	tRNA
289086	1.51	1	13	24	0	21.6	0.39	Unknown

involvement in the processing of this transcript. In another example, the remaining two groups of sequences aligned to the chloroplast polycistronic transcript that includes *psbB*, *psbT*, *psbH*, *petB*, and *petD* in two separate regions. Multiple examples of each of these transcripts have been inserted into the *O. sativa* chromosome, and it is impossible to determine whether these sequences derive from their nuclear or chloroplast copies. However, as the small RNA processing machinery is generally thought to reside in the nucleus, it is likely that these small RNAs are of nuclear origin.

Discussion

In-depth analysis of small RNAs from organisms at strategically chosen phylogenetic distances is necessary to gain a better understanding of the evolution of plant cellular process mediated by small RNA. However, to date, few efforts have focused on small RNAs in any plant species outside of the angiosperms. In this global survey of small RNAs in *P. contorta*, many gymnosperm miRNAs with known angiosperm homologs have been identified (Supplemental Table 1). Though many of the so-called conserved miRNAs have been found in a variety of plant species, only direct sequencing of the small RNA molecules provides a definitive route to discover novel miRNAs outside these families. With our introduction of a novel sequence-based clustering method and a support vector machine that classifies the resultant clusters, this study has also provided an unprecedented view of the miRNAs present in *P. contorta*, a species with limited expressed sequence and no genomic sequence data.

Owing to our application of the miRNA annotation techniques developed in this work, these data have not been limited to the identification of only those miRNAs with homologs in *A. thaliana* or other plants (see Tables 1 and 2). Two methods for identifying novel miRNAs were applied to the *P. contorta* small RNA sequences. One of these methods considers putative pre-miRNA structure, and the other relies on identification of miRNA-like sequence clusters. These techniques have revealed a set of likely gymnosperm-specific novel miRNAs, a few novel *O. sativa* miRNAs, as well as two novel miRNAs that appear to be conserved between *O. sativa* and *P. contorta* (Table 1).

Though many of their pre-miRNA structures have not been validated here, the high expression of many of these predicted miRNAs as well as biases toward 5′ terminal uridine and 3′ terminal guanosine residues provides secondary support that these sequences are miRNA-like. Many (19) of the *P. contorta* small RNA clusters summarized in Table 2 had at least one small RNA sequence with a perfect alignment to an EST sequence, but only five of these folded into an acceptable fold-back structure. Though this may appear to reject these remaining candidates as true miRNAs, this is not necessarily the case. The mRNA targets of most plant miRNAs have perfect or near-perfect recognition sites. The observation that a small RNA sequence has perfect alignments to multiple unrelated mRNAs implicates those mRNAs as potential target transcripts. To support this, putative *P. contorta* target transcripts for all miRNA sequences in Tables 1 and 2 were predicted using miRU (Zhang 2005). If one method for miRNA identification performed better, one might expect a difference in the number of predicted targets for these miRNAs owing to a lack of complementarity of non-miRNA sequences within the transcriptome. The number of predicted target genes did not differ significantly between the miRNAs in Tables 1 and 2, supporting that our novel SVM method is comparable to the standard folding-based method for novel miRNA identification ($P = 0.3409$, Kruskal–Wallis rank-sum test). Still, the predicted novel miRNAs presented in Table 2 should be approached with caution until their pre-miRNA structure can be validated by another approach. Furthermore, even if some of these small RNA molecules cannot presently be defined as miRNAs by current standards, their high expression and reproducible endonuclease cleavage positions support that they perform an important function in *P. contorta*, perhaps acting as *trans*-acting siRNAs or potentially in some other uncharacterized processes. That the putative miRNAs provided here were identified by a signature of precise and reproducible maturation suggests that they are processed in a pathway separate from the apparent randomly-derived degradation products of rRNA and tRNA that were observed in high numbers in this and similar studies.

Apart from known *O. sativa* miRNAs and novel gymnosperm miRNAs, other groups of small RNAs were identified that are highly conserved between these two species. The *O. sativa* 24-nt sequences showed evidence for heterochromatin siRNA activity as their predicted target sites corresponded to centromeric regions and rRNA gene clusters (Fig. 3). Though they contained a relatively low proportion of 24-nt small RNAs, those 24-nt sequences found in *P. contorta* were more apt to align to the *O. sativa* genome than 23-nt or 25-nt small RNAs (Fig. 2A,E). However, very few of the *O. sativa* 24-nt heterochromatin siRNAs have identifiable *P. contorta* homologs based on sequence clustering. Also, those that do appear to have homologs are not strictly 24 nt in length, and many correspond to full-length or partial rRNA genes rather than transposable elements. Taken together, it appears that the pathways producing heterochromatin siRNAs existed prior to the divergence of these species. Though the length difference of supposed siRNAs between *P. contorta* and *O. sativa* suggests that this class of siRNAs may derive from related genomic repeats, their exact mode of maturation may differ in the gymnosperms. A larger population of siRNAs responsible for controlling transposable elements would be expected in the gymnosperms since much of the difference in genome size between these plants is posited to be due to expansion of such entities (Bennett and Leitch 2005). The larger diversity of 21-nt RNAs in *P. contorta* is consistent with the hypothesis that ~21-nt small RNAs are fulfilling a substantial portion of this role.

A few other interesting groups of small RNAs that are conserved between *O. sativa* and *P. contorta* were found here. Small RNAs from a few of these groups matched only to genes from the chloroplast (or chloroplast-derived genes that have inserted into the nuclear genome). Some of these small RNA sequences have also been observed in similar studies of *O. sativa* (Chen et al. 2006) and *Arabidopsis* (Lu and Tei 2005; Rajagopalan et al. 2006). It is unclear whether these sequences are derived from these transcripts or target them for enzyme processing (or both). Two of these sequences map ~150 nt upstream of a number of chloroplast tRNA genes or their nuclear-encoded counterparts. The other three groups aligned to distinct regions of two polycistronic chloroplast transcripts. The site between the two genes on the *ndhB*/*rps7* transcript is of particular interest because it is centered across the experimentally determined endonuclease cleavage site (Hashimoto et al. 2003) (Supplemental Fig. 5b). One of the two small RNA alignment sites within the polycistron containing *psbB, psbT, psbH, petB*, and *petD* has previously been identified, using comparative genomics, as a potential recognition site of a protein regulating mRNA processing (Seliverstov and

Lyubetsky 2006). These data suggest that this site is conserved not for recognition by a protein but rather as a small RNA binding site that effects cleavage of the polycistron, perhaps by the enzyme(s) responsible for degrading transcripts targeted by the miRNA pathway. This may suggest another function of the miRNA pathway and, considering that it appears to be limited to chloroplast-derived transcripts and is deeply conserved, it may reflect a more ancient application of small RNA-directed transcript processing.

Comparison of the small RNAs expressed in *P. contorta* to *O. sativa* has revealed known and novel conserved miRNAs. As well, alignment of the *P. contorta* sequences to the *O. sativa* genome identifies a set of perfectly conserved small RNAs, likely ancient and diverse in origin and function. This result supports that miRNA and siRNA pathways were likely functional in the common ancestor of these two species. We also conclude, based on our discovery of up to 53 novel miRNA families in *P. contorta* (12 in Table 1 and 44 in Table 2, with 1 shared) and a plethora of remaining un-annotated small RNAs, that the diversity of small RNA-mediated processes in the gymnosperms when combined with that of the angiosperms, will provide important context for understanding the evolution of RNA silencing in the spermatophytes.

Methods

Small RNA isolation (*P. contorta*)

Approximately 0.5 g of young needles was collected and ground into fine powder with a pestle and mortar in the presence of liquid nitrogen. The powder was transferred into a precooled tube and suspended quickly in 1 mL of RNA extraction buffer (100 mM LiCl, 1% SDS, 10 mM EDTA, 100 mM Tris at pH 9); ~1.5 mL of warm phenol was added, and after the tube cooled down to room temperature, ~1.5 mL of chloroform was added, and the mixture was extracted by inverting the tube for 20 min. The tube was centrifuged, and the supernatant was transferred to a fresh tube and extracted twice again in a 50:50 mixture of phenol/chloroform. Nucleic acids were precipitated by the addition of 0.1 volume of 3 M sodium acetate (pH 5) and 2 volumes of 100% ethanol (relative to the extracted sample volume), pelleted by centrifugation, air-dried, and dissolved in 0.2 mL of water. To precipitate small RNA molecules, the nucleic acid solution was mixed with NaCl at the final concentration of 300 mM, glycogen at the final concentration of 12 µg/mL, and 2.5 volumes of 100% ethanol. Nucleic acids were ultimately precipitated at −20°C overnight, which was followed by centrifugation at 4°C for 30 min at 13,200*g*. *O. sativa* small RNAs were extracted using the same protocol using leaf buds as a starting material.

Small RNA pool preparation

Small RNA from *O. sativa* and *P. contorta*, samples with gel mobility in the 15- to 30-nt size range were gel purified and ligated to an adenylated DNA 3′ adapter with the sequence 5′-AppGAAGAGCCTACGACGA (adapter at 20 µM, 50 mM HEPES, pH 8.3, 10 mM $MgCl_2$, 3.3 mM DTT, 10 µg/mL BSA, 8.3% glycerol) using 4 U/µL T4 RNA ligase (GE Amersham Biosciences) for 90 min at room temperature. The resulting RNA–DNA hybrids were gel purified using 10% PAGE and ligated to a 5′ RNA adaptor having the sequence 5′-rAUCGUAGGCACCUGAAA. This ligation utilized the adaptor at 30 µM. The resulting material was ethanol precipitated and reverse transcribed using a long primer containing sequence required for 454 Sequencing together with a 10-nt random sequence element, 5′-CCTATCCCCTGTGTGCCTTGCCTATCCCCTGTTGCGTGTCTCAG$(N)_{10}$TCGTCGTAGGCTCTTC. Reverse transcription was with SuperScript II (Invitrogen) using the supplied protocol. The RNA template was destroyed by heating in the presence of 100 mM KOH, and the resulting cDNA was isolated on a 10% denaturing polyacrylamide gel. PCR was conducted using a biotinylated primer (5′-BCCTACCCCTGTGTGCCTTGCCTATCCCCTGTTGCGTGTCTCAG, B indicates location of Biotin residue) and 5′-primers (*O. sativa*, 5′-CCATCTCATCCCTGCGTGTCCCATCTGTTCCCTCCCTGTCTCAGTGCTAAGCATCGTAGGCACCTGAAA; *P. contorta*, 5′-CCATCTCATCCCTGCGTGTCCCATCTGTTCCCTCCCTGTCTCAGTAGATACGATCGTAGGCACCTGAAA). Both primers were used at a final concentration of 0.5 µM.

Northern blot detection of miRNAs in *P. contorta*

RNA was extracted from the green needles of cold acclimatized *P. contorta* seedlings. Northern blot was produced by running 20 µg of RNA into a 15% denaturing polyacrylamide gel and transferring to Hybond-N+ (Amersham) membrane using a NovaBlot (Pharmacia) electrophoresis unit. The gel was stained with SYBR Green II to visualize a 21-nt size standard. All other steps in the protocol were according to those previously described (Lau et al. 2001). DNA probes were 5′-AAGTTCAAGAAAGCTGTGGA (for miR396), 5′-AACCATCGAGGACCTGACGT (for the first novel miRNA, likely homologous to miR950b), and 5′-TTGGGAAA TAGTTGATAATA (for the novel miRNA with no EST support).

Small RNA sequencing and sequence processing

Small RNA samples were sequenced using the high-throughput pyrosequencing developed by 454 Life Sciences (454 Life Sciences, GS20 platform) (Margulies et al. 2005). Each sequence read is thought to represent a single adaptor-ligated small RNA molecule. The reads were searched for adaptor sequences and library identification tags. The small RNA sequences were assumed to be the sequence between the 5′ and 3′ adaptor sequences. Intervening sequences 10 nt or shorter were ignored. The random 10-nt sequence was assumed to be the first 10 bases following the 3′ adaptor sequence. All sequences were extracted and stored in our custom-built database named MyRNA, which will be made available for public use in the near future. Upon encountering the same small RNA sequence more than once, the sequence count was incremented only if the 10-nt-long random sequence tag was unique. For sequence comparisons, the MyRNA database was also loaded with all plant mature miRNA sequences from miRBase release 9.1 (http://microrna.sanger.ac.uk), which were included in the sequence-clustering pipeline (see below).

Partitioning small RNAs into clusters of related sequences

Small RNA sequences in the MyRNA database were compared in an all-against-all comparison using a heuristic implementation of the Needleman–Wunsch global alignment algorithm (Needleman and Wunsch 1970). Rather than producing alignments, this implementation returns an approximate edit distance between two sequences. As sequences differing in length would be missed by this approach, the algorithm was modified to perform global alignments between all equal-length subsequences of the longer sequence to the shorter sequence of a given pair. Only alignments of 19-nt or longer were kept, since shorter alignments increased the clustering of unrelated sequences. Edit distances of less than four in any alignment were deemed significant and stored in the database. The sequences were considered nodes, and the calculated edit distances were used to weight the edges of the graph. All connected components were extracted from this graph, revealing highly similar groupings of small RNA se-

quences within the MyRNA database. This method mimics the use of single-linkage hierarchical clustering for three iterations (Gower and Ross 1969). Sequences within each cluster were aligned using ClustalW. The information content was calculated at each position of the alignment and averaged across the total length of the consensus (Schneider and Stephens 1990).

Genome mapping and small RNA annotation

Sequences were mapped to the *O. sativa* genome using MegaBlast (Zhang et al. 2000) with low-complexity filtering disabled. Only alignments with perfect identity across the length of the query sequence were stored in the MyRNA database. *P. contorta* small RNA sequences were aligned both to the *O. sativa* genome and to the set of *P. taeda* ESTs and cDNAs available from GenBank as of July 2006. Perfect matches of *P. contorta* small RNA sequences to *P. taeda* ESTs were also stored in the database.

O. sativa genome annotations (build 3) were downloaded from the RAP1 website (http://rapdownload.lab.nig.ac.jp/index.html). Ribosomal RNA gene coordinates were obtained separately from this group. Coordinates of genomic repeats were determined using RepeatMasker (www.repeatmasker.org), RepBase version 9.04. The known positions of *O. sativa* miRNA genes were obtained from miRBase. All annotations and small RNA sequence positions were loaded into a local database and accessed through our instance of the GBrowse genome browser (Stein et al. 2002), available at our website (http://microrna.bcgsc.ca/cgi-bin/gbrowse/rice_build_3/).

All small RNA sequences with known genomic positions were annotated based on any overlap with these annotations. Because a small RNA sequence could have more than one genomic position, and thus could overlap with multiple annotations, a priority was assigned to each annotation type. The highest priority was given to overlap with known miRNA genes, followed by overlap with rRNA and tRNA genes. Any un-annotated small RNAs were then annotated as "genomic repeat" if they overlapped at least one RepeatMasker annotation. The remaining small RNAs were further annotated by BLASTN search against Rfam (release 8.0) using an *E*-value threshold of 0.01. Sequences with no hits in Rfam were labeled "un-annotated" in the database.

There were many sequences that did not map to the *O. sativa* genome perfectly or were still classified as "un-annotated" after the positional annotation process. Further annotation of these sequences was accomplished by employing the small RNA sequence-based clusters described above. In general, all sequences in the same cluster share high sequence identity. The appropriate annotation was first added to small RNAs residing in a cluster that contained annotated sequences (from above) following the same priority to annotations as before. Sequences in clusters containing more than 100 sequences were ignored in this process, as these clusters were generally low in information content.

Degenerate mapping of *O. sativa* siRNAs to genomic sequence

The modified local alignment algorithm described in the clustering section was altered to allow searches against genomic sequence. Alignments were kept if they were no shorter than $L - 3$, where L is the query sequence length. No more than a total of three mismatches or insertions/deletions were allowed within the aligned region.

Discovery of nonconserved *P. contorta* miRNA families using ESTs

Owing to the lack of genomic sequence, annotation of small RNAs as miRNAs could only employ publicly available ESTs and cDNA sequence as well as miRNAs from other plant species. Because most of the pre-miRNA sequences are rapidly degraded after DCL cleavage, only the mature miRNA and sometimes the miRNA* sequence is obtained. Using this information, the small RNAs with cDNA/EST alignments were checked for miRNA-like alignment patterns. An EST/cDNA with miRNA-like alignment was defined here as having no more than three clusters of small RNA sequences on it, with each cluster of length no larger than 30-nt. Folding these sequences with a standard free energy minimization folding algorithm (Hofacker 2003) facilitated an improved method of true miRNA identification. It is known that plant pre-miRNAs vary from ~80 to ~160 nt in length (Zhang et al. 2006b). Rather than folding the entire EST sequence, only a region of 150 nt on either side of the sequence clusters was folded. Of these folded flanking sequences, the one with the lower MFE was considered the putative pre-miRNA structure. The MFE of the folded sequences was considered as an additional support for the potential of many of these sequences to form stable pre-miRNA structures. Sequences producing structures with MFE > -25 kcal/mol were ignored as well as those with less than half the nucleotides of the small RNA nucleotides paired in the most stable structure.

SVM-aided discovery of novel nonconserved *P. contorta* miRNAs

Trends of information content and cluster size suggested that clusters of miRNAs might be distinguished based on various parameters derived from the clusters. A SVM was employed as a secondary approach to classify small RNA clusters that may lack EST sequences. The support vector machine implementation entitled SMO was used in the Weka software package (Witten and Frank 2005). The positive training data comprised all clusters containing less than 100 sequences and at least one known (conserved) miRNA sequence. Negative training examples were those clusters that could be classified as other types of small RNAs (tRNA, repeat-derived or rRNA). The following traits were computed for each cluster as parameters for the classifier: number of sequences from *P. contorta* and *O. sativa*, mean information content, number of genomic loci (rice genome), mean length of sequences in the cluster, and the variance of these lengths. Suitable parameters were chosen by applying 10-fold cross validation to the training set. The non-default parameters used for prediction were as follows: exponent = 3, $c = 1.5$. When the classifier was trained as described, it could classify clusters of known miRNAs with modest sensitivity (0.645) but high specificity (0.952). This suggested it was a reliable method for the prediction of putative novel miRNA clusters from these data with a very low rate of false-positive predictions.

Acknowledgments

The *P. contorta* samples were provided by Jim Mattsson, Department of Biology, Simon Fraser University, Canada. Rice small RNA extracts were a generous gift from M.B. Wang. This project was funded in part by the Natural Sciences and Engineering Research Council of Canada (H.A.E. and P.J.U.). We thank Marco Marra of the BC Genome Sciences Centre for critical evaluation of the manuscript. P.J.U. is a Senior Michael Smith scholar. S.C.S. is a Michael Smith scholar and a Canadian Research Chair. R.D.M. and G.A. receive stipends from the Canadian Institutes for Health Research and the Michael Smith Foundation for Health Research.

References

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., et al.

2003. A uniform system for microRNA annotation. *RNA* **9:** 277–279.
Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.
Bennett, M.D. and Leitch, I.J. 2005. Nuclear DNA amounts in angiosperms: Progress, problems and prospects. *Ann. Bot.* **95:** 45–90.
Bonnet, E., Wuyts, J., Rouze, P., and de Peer, Y.V. 2004. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc. Natl. Acad. Sci.* **101:** 11511–11516.
Borsani, O., Zhu, J., Verslues, P.E., Sunkar, R., and Zhu, J.K. 2005. Endogenous siRNAs derived from a pair of natural *cis*-antisense transcripts regulate salt tolerance in *Arabidopsis*. *Cell* **123:** 1279–1291.
Chen, Z., Zhang, J., Kong, J., Li, S., Fu, Y., Li, S., and Zhang, H. 2006. Diversity of endogenous small non-coding RNAs in *Oryza sativa*. *Genetica* **128:** 21–31.
Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R., and Jiang, J. 2002. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14:** 1691–1704.
Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L., et al. 2007. High-throughput sequencing of *Arabidopsis* microRNAs: Evidence for frequent birth and death of miRNA genes. *PLoS ONE* **2:** e219. doi: 10.1371/journal.pone.0000219.
Gower, J.C. and Ross, G.J.S. 1969. Minimum spanning trees and single linkage cluster analysis. *Appl. Stat.* **18:** 54–64.
Griffiths-Jones, S. 2006. miRBase: The microRNA sequence database. *Methods Mol. Biol.* **342:** 129–138.
Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. 2003. Rfam: An RNA family database. *Nucleic Acids Res.* **31:** 439–441.
Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C., and Kasschau, K.D. 2005. ASRP: The *Arabidopsis* Small RNA Project Database. *Nucleic Acids Res.* **33:** D637–D640.
Hashimoto, M., Endo, T., Peltier, G., Tasaka, M., and Shikanai, T. 2003. A nucleus-encoded factor, CRR2, is essential for the expression of chloroplast *ndhB* in *Arabidopsis*. *Plant J.* **36:** 541–549.
Herr, A.J. 2005. Pathways through the small RNA world of plants. *FEBS Lett.* **579:** 5879–5888.
Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31:** 3429–3431.
Itoh, T., Tanaka, T., Barrero, R.A., Yamasaki, C., Fujii, Y., Hilton, P.B., Antonio, B.A., Aono, H., Apweiler, R., Bruskiewich, R., et al. 2007. Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*. *Genome Res.* **17:** 175–183.
Johnson, C., Bowman, L., Adai, A.T., Vance, V., and Sundaresan, V. 2007. CSRDB: A small RNA integrated database and browser resource for cereals. *Nucleic Acids Res.* **35:** D829–D833.
Jones-Rhoades, M.W. and Bartel, D.P. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* **14:** 787–799.
Lau, L.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* **5543:** 858–862.
Lindow, M. and Krogh, A. 2005. Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics* **6:** 119. doi: 10.1186/1471-2164-6-119.
Lu, C. and Tei, S.S. 2005. Elucidation of the small RNA component of the transcriptome. *Science* **309:** 1567–1569.
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.
McCarthy, E.M., Liu, J., Lizhi, G., and McDonald, J.F. 2002. Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biol.* **3:** RESEARCH0053. doi: 10.1186/gb-2002-3-10-research0053.
Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., et al. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* (this issue). doi: 10.1101/gr.7179508.
Nakano, M., Nobuta, K., Vemaraju, K., Tej, S.S., Skogen, J.W., and Meyers, B.C. 2006. Plant MPSS databases: Signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res.* **34:** D731–D735.
Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.
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.
Pavy, N., Paule, C., Parsons, L., Crow, J.A., Morency, M.J., Cooke, J., Johnson, J.E., Noumen, E., Guillet-Claude, C., Butterfield, Y., et al. 2005. Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. *BMC Genomics* **6:** 144.
Pontier, D., Yahubyan, G., Vega, D., Bulski, A., Saez-Vasquez, J., Hakimi, M.A., Lerbs-Mache, S., Colot, V., and Lagrange, T. 2005. Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in *Arabidopsis*. *Genes & Dev.* **19:** 2030–2040.
Pontius, J.U., Wagner, L., and Schuler, G.D. 2002. UniGene: A unified view of the transcriptome. In *The NCBI handbook*. National Library of Medicine, Bethesda, MD.
Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., and White, J. 2001. The TIGR gene indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29:** 159–164.
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.
Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18:** 6097–6100.
Seliverstov, A. and Lyubetsky, V. 2006. Translation regulation of intron-containing genes in chloroplasts. *J. Bioinform. Comput. Biol.* **4:** 783–792.
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.
Talmor-Neiman, M., Stav, R., Klipcan, L., Buxdorf, K., Baulcombe, D.C., and Arazi, T. 2006. Identification of *trans*-acting siRNAs in moss and an RNA-dependent RNA polymerase required for their biogenesis. *Plant J.* **48:** 511–521.
Vazquez, F. 2006. *Arabidopsis* endogenous small RNAs: Highways and byways. *Trends Plant Sci.* **11:** 460–468.
Wang, M.B. and Metzlaff, M. 2005. RNA silencing and antiviral defense in plants. *Curr. Opin. Plant Biol.* **8:** 216–222.
Williams, L., Carles, C.C., Osmont, K.S., and Fletcher, J.C. 2005. A database analysis method identifies an endogenous *trans*-acting short-interfering RNA that targets the *Arabidopsis ARF2*, *ARF3*, and *ARF4* genes. *Proc. Natl. Acad. Sci.* **102:** 9703–9708.
Witten, I.H. and Frank, E. 2005. *Data mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, San Francisco, CA.
Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* **2:** e104. doi: 10.1371/journal.pbio.0020104.
Yao, Y., Guo, G., Ni, Z., Sunkar, R., Du, J., Zhu, J.K., and Sun, Q. 2007. Cloning and characterization of microRNAs from wheat (*Triticum aestivum* L.). *Genome Biol.* **8:** R96. doi: 10.1186/gb-2007-8-6-r96.
Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P., and Anderson, T.A. 2006a. Conservation and divergence of plant microRNA genes. *Plant J.* **46:** 243–259.
Zhang, B., Pan, X.P., Cox, S.B., Cobb, G.P., and Anderson, T.A. 2006b. Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.* **63:** 246–254.
Zhang, Y. 2005. miRU: An automated plant miRNA target prediction server. *Nucleic Acids Res.* **33:** W701–W704.
Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7:** 203–214.

Received July 16, 2007; accepted in revised form December 17, 2007.