Published in final edited form as:

Curr Opin Plant Biol. 2014 April ; 0: 87–95. doi:10.1016/j.pbi.2014.02.008.

Seeing the forest for the trees: Annotating small RNA producing genes in plants

Ceyda Coruh^{a,b,c}, Saima Shahid^{a,b,c}, and Michael J. Axtell^{a,b,c}

^aDepartment of Biology, Penn State University, University Park, PA 16802 USA

^bPlant Biology Intercollegiate Ph.D. Program, Penn State University, University Park, PA 16802 USA

^cHuck Institutes of the Life Sciences, Penn State University, University Park, PA 16802 USA

Abstract

A key goal in genomics is the complete annotation of the expressed regions of the genome. In plants, substantial portions of the genome make regulatory small RNAs produced by Dicer-Like (DCL) proteins and utilized by Argonaute (AGO) proteins. These include miRNAs and various types of endogenous siRNAs. Small RNA-seq, enabled by cheap and fast DNA sequencing, has produced an enormous volume of data on plant miRNA and siRNA expression in recent years. In this review, we discuss recent progress in using small RNA-seq data to produce stable and reliable annotations of miRNA and siRNA genes in plants. In addition, we highlight key goals for the future of small RNA gene annotation in plants.

Introduction

In plants, a particularly wide variety of small regulatory RNAs are produced by Dicer-Like proteins (DCLs) and utilized as sequence-specific guides by Argonaute (AGO) proteins. The known DCL/AGO-associated small RNAs (hereafter, small RNAs) are 20-24 nts in length. They function to guide repressive regulation of targets selected on the basis of small RNA-target complementarity at the transcriptional or post-transcriptional levels. Several major types have been described, including miRNAs, secondary short interfering RNAs (secondary siRNAs), and heterochromatic siRNAs (Fig. 1). Huge amounts of small RNA alignment data have been produced using small RNA-seq, and progress at using these alignments to create small RNA gene annotations has been made.

© 2014 Elsevier Ltd. All rights reserved.

Author for correspondence: Michael J. Axtell (mja18@psu.edu). PH: (+01) 814-867-0241.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Complications in microRNA annotation

*MIRNA*s (the loci which produce mature miRNAs) have received much attention and are thus the most well-annotated type of small RNA genes in plants. *MIRNA* annotations are disseminated by miRBase [1]. Currently, miRBase (release 20) houses annotations of hundreds of *MIRNA* genes from 72 plant species. Community accepted standards specific for the features of plant *MIRNA*s guide miRBase submissions [2]. The basic premise of miRBase is that a hairpin RNA transcribed from the *MIRNA* locus is processed to ultimately yield a single functional mature miRNA; the minimal miRBase entry consists simply of a hairpin and a single linked mature miRNA sequence. However, the reality of miRNA expression is now known to be much more complex.

Related *MIRNA* hairpins often produce mature miRNAs that vary in length, sequence, or both. This variation can result from expression of multiple paralogous *MIRNA*s that differ slightly in sequence, creating several slightly different mature miRNAs. Another, very common type of miRNA variation is the result of differentially processed and/or truncated RNAs from the same hairpin (Fig. 2A). To illustrate how common such variation is, we aligned small RNA-seq data from wild-type *Arabidopsis* flowers and leaves (NCBI GEO GSM738731 and GSM738727; [3]) to the *Arabidopsis* nuclear genome, and compared the alignments to annotations from miRBase 20. Precision^{ann} values (the fraction of all alignments to a hairpin corresponding to the miRBase-annotated mature miRNA) were often very poor (Fig. 2B). The distribution of precision^{max} values (the fraction of all alignments to a hairpin corresponding to the most abundantly observed small RNA) values was better, but nonetheless showed that it is very rare for an annotated *MIRNA* hairpin to produce just one discrete RNA (Fig. 2C). In our analysis the most abundant RNA was NOT annotated as the mature miRNA for the majority of *Arabidopsis* *MIRNA* loci (Fig. 2D). According to our current understanding, only AGO-loaded small RNAs are functional. There is no guarantee that all RNAs observed via small RNA-seq are AGO-bound. We therefore aligned a set of small RNAs that co-immunoprecipitated with a major *Arabidopsis* AGO protein, AGO1 (NCBI GEO GSM989351; [4]), and performed a similar analysis. Based on the known preferences of AGO1 for RNA binding, this analysis was limited to *MIRNA* loci whose annotated mature miRNAs were 21 nts with a 5'-U. The distributions of precision values improved (Figs. 2E-F), as did the concordance between miRBase annotations and the observed most abundant RNAs (Fig. 2G). Nonetheless, extensive heterogeneity in miRNA accumulation was still apparent for nearly all known *MIRNA* loci. Two conclusions emerge from this simple analysis. One: there are large discrepancies between empirical data and miRBase in terms of annotation of the mature miRNA. Two: even putting aside potential errors in annotation of mature miRNAs, nearly all known *MIRNA* hairpins produce more than a single product.

One type of alternative RNA that arises from *MIRNA* hairpins are miRNA*s. In the canonical viewpoint of miRNA biogenesis, the miRNA* is defined as the strand of the initial miRNA/miRNA* duplex that is discarded at the time of AGO-loading. However, there is ample evidence demonstrating that miRNA*s can also be AGO-loaded and functional. Many miRNA*s are enriched in AGO1 immunoprecipitates [5], others associate with AGO2 [6], and several have known functions [6-7]. Positional variants outside of the

annotated miRNA/miRNA* pair are also prominent features of plant *MIRNA* hairpin processing and they are known to have functional consequences [8]. A very extensive study by Jeong et al. [9**] demonstrated that heterogeneity in *MIRNA* processing is quite common in *Arabidopsis*, and that in many cases there is compelling evidence for the functional relevance of these processing variants.

Additional complexity in miRNA annotation arises due to various modifications of mature miRNAs that occur after dicing. HEN1 is a methyltransferase that catalyzes 2'-O-methylation of the 3'-most nucleotide of plant miRNAs and siRNAs [10]. In *hen1* mutants, miRNAs display extensive 3'-truncations coupled with addition of non-templated nts (predominantly U) at the 3' end [11**]. The truncated and tailed variants occur after the miRNAs are loaded onto the AGO1 protein, implying that these modifications could potentially affect the target specificity of the miRNAs. Importantly, 3'-truncation and 3' non-templated tailing also occur for some miRNAs in the wild-type background [11**], implying that this may be a mechanism used in normal conditions to modulate miRNA target specificity or mechanism of action.

MIRNA Superfamilies

Another challenge in miRNA annotation is to accurately describe the evolutionary relationships between *MIRNA* loci. *MIRNA* loci are commonly grouped into families (which are assigned the same number) based on high levels of sequence similarity. However, the existence of *MIRNA* superfamilies, whose members have evidence of common descent and functions despite extensive sequence diversification, complicates this system. In one extreme example, both *Physcomitrella patens* (a moss) and flowering plants express miRNAs (miR904 and miR168, respectively) that target *AGO1* mRNAs, but the mature miRNAs have no detectable sequence similarity [12]. Whether this situation arose because of convergent evolution or extensive sequence diversification of a single ancestral miRNA is not clear. The miR482/2118 superfamily of miRNAs comprise a sequence-diverse set of mature miRNAs that are present in many plant species, and frequently function to target nucleotide binding site-leucine-rich repeat (*NB-LRR*) innate immune receptor mRNAs [13*, 14**, 15*], as well as other RNAs [16]. A second set of plant *MIRNA* superfamilies is comprised of the miR390, miR4376, and miR7122 superfamilies [17**]. Members of the miR390 superfamily are highly conserved in most plant species, but miR4376 and miR7122 superfamilies have highly diverse mature miRNAs in various species. Careful sequence analysis provides compelling evidence that the miR390, miR4376, and miR7122 superfamilies are all related by common descent [17**]. Curiously, all of these described superfamilies serve as initiators of secondary siRNA biogenesis. The observation of superfamilies whose members have diverged to the edge of reliable alignments suggests that many other evolutionary relationships between superficially unrelated *MIRNA*s may exist.

The annotation gap

miRBase is the main source for *MIRNA* annotations for all organisms. However, it is critical to emphasize just how minor the contribution of miRNAs are to the total small RNA expression profile of plants. To illustrate this, we compared *Arabidopsis* small RNA-seq

alignments to miRBase annotations. As a counterpoint, we also compared aligned polyA+ RNA-seq data to the TAIR10 mRNA annotations. The small RNA-seq dataset was from flowers and leaves as used in Figure 2 [3]. The RNA-seq dataset was also derived from flowers and leaves, and comprised 101 nt single-end reads from polyA-enriched samples [18]. To minimize contamination with breakdown products of abundant RNAs, rRNA, tRNA, snRNA, and snoRNA regions of the reference genome sequence were masked prior to alignment, and only alignments of 20-24 nt reads were retained for the small RNA-seq data. The RNA-seq data were aligned using a spliced aligner (tophat; [19]) and randomly down-sampled to achieve an approximately equal number of alignments compared to the small RNA-seq data (32.5E6 and 35.7E6 alignments for the small RNA-seq and RNA-seq data, respectively). For the purposes of illustration, we considered a genomic position active if it had a coverage ≥ 0.1 reads per million, which equated to a depth of four or more alignments for both datasets. Based on this analysis, roughly 34 million and 12 million nucleotides of the *Arabidopsis* genome expressed significant polyA+ and small RNA, respectively (Fig. 3A). There was very little overlap between the two, indicating that expression of long polyA+ RNA and 20-24 nt RNAs is usually mutually exclusive in these tissues. Annotated *MIRNA* loci account for only a tiny fraction of the genome that actively produces 20-24 nt RNAs (Fig. 3B, left). In contrast, nearly all of the polyA+ RNA-seq is explained by existing gene annotations (TAIR10; Fig. 3B, right). In terms of abundance, small RNAs aligned to annotated *MIRNA* hairpins were in the minority; however, nearly all of the polyA+ RNA-seq alignments fell within annotated genes (Fig. 3C). We do not believe this analysis implies a vast amount of un-annotated *MIRNA* loci. Instead, it highlights the fact that the majority of expressed plant small RNAs are NOT miRNAs, and that these in total account for roughly 10% of the *Arabidopsis* genome. Clearly, there is large ‘annotation gap’ between the empirical knowledge of small RNA expression and the annotations of small RNAs provided by miRBase alone.

Other hairpin-derived RNAs

Long inverted-repeat containing hairpin RNAs (hpRNAs) have long been used to manipulate plant mRNA expression levels [20,21]. Small RNAs derived from artificial hpRNA constructs are processed in a manner similar to the processing of viral RNAs and drive silencing of endogenous and exogenous genes as well as trigger long distance signals in *Arabidopsis* [22]. Genome-wide scans find substantial correlations between small RNA accumulation and hairpins that do not qualify as miRNAs [23,24*], implying that endogenous hpRNAs may be widespread. Only a few endogenous hpRNA loci have been characterized in depth. These include the *IR71* and *IR2039* loci in *Arabidopsis* [25] and the *Mu killer* locus in maize [26]. Systematic annotation of endogenous hpRNA loci has not yet been reported, and there are not yet clear community-accepted standards for discerning hpRNA loci. Nonetheless, the available evidence suggest that there may be a great number of such genes.

Secondary, phased siRNAs

Secondary siRNAs are characterized by a distinct small RNA biogenesis pathway which requires the slicing of a primary transcript by a specific miRNA or other secondary siRNAs.

The cleaved transcript is converted into a dsRNA by an RNA-dependent RNA polymerase and then processed by a DCL protein into siRNAs [27]. Because the location of the initial cut is specified by an upstream small RNA cleavage, dicing of the dsRNA with a defined start point generates siRNAs in a “phased” pattern. Most annotated secondary siRNAs have been found using several similar algorithms based upon this characteristic phased pattern [28,29]. However, in contrast to *MIRNA* loci, there is as yet no centralized database or registry devoted to this class of small RNA loci.

The classic examples of phased secondary siRNA loci are several families of non protein-coding RNAs termed *TRANS ACTING siRNA (TAS)* loci. Some phased siRNAs can repress target mRNAs in *trans*, hence the term *trans*-acting siRNAs (tasiRNAs). The extensively-studied *TAS3a/b/c* family is targeted at two sites by miR390 and produces conserved tasiRNAs that target *Auxin Response Factor (ARF)* mRNAs involved in developmental timing and leaf polarity [30]. *TAS3* is a particularly well-conserved *TAS* locus, and even has homologs in the moss *Physcomitrella patens* [31]. A linkage between miR390-controlled *TAS3* loci and a novel miR156-controlled *TAS* family, *TAS6*, has been identified in *Physcomitrella* [32-33*]. *TAS6a* and *TAS3a* are present on the same primary transcript, which has four target sites: two for miR156, that define the *TAS6a* region, and two for miR390, which define the *TAS3a* region. Inhibition of miR156 and over-expression of miR390 both delayed gametophore development, and resulted in the increased production of miR390-triggered tasiRNAs [33*]. These data demonstrate that *TAS* transcripts can serve as integration points that sense and respond to the accumulation of multiple miRNAs.

Protein-coding genes also can spawn secondary, phased siRNAs. Phased siRNAs from diverse sets of protein-coding genes have been observed in multiple plant species (reviewed by [27]). Assuming that some of the induced secondary siRNAs can act as tasiRNAs to target other members of large gene families, secondary phased siRNA production from protein-coding mRNAs may serve as a mechanism to achieve coordinate post-transcriptional repression for many transcripts at once. One example of special interest is the miR482/2118 superfamily, which target *NB-LRR* disease resistance mRNAs. In *Medicago truncatula*, miR2118, miR2109, and miR1507 cause large amounts of phased secondary siRNAs from at least 71 *NB-LRR* mRNAs [13*]. High accumulation of these three miRNAs is seen across the Fabaceae [13*]. In tobacco, miR6019 and miR6020 target the *N* resistance gene and cause extensive production of secondary phased siRNAs [15*]. In tomato, sequence diverse members of the miR482 family also target large numbers of *NB-LRR* mRNAs, which in turn produce phased siRNAs [14**]. Importantly, both viral and bacterial infections of tomato correlate with decreased miR482 accumulation and increased *NB-LRR* accumulation [14**]. This suggests that pathogen-induced suppression of miRNA levels could serve to enhance *NB-LRR* expression, perhaps priming plant defense responses. This has the potential to be a wide-spread mechanism, as *NB-LRR* mRNAs are potent sources of phased siRNAs in many plant species, including the conifer *Picea abies* [34].

Heterochromatic siRNAs

Heterochromatic siRNAs are the major components of the small RNA populations in most tissues of most plant species examined to date. Most angiosperm genomes have thousands of

loci that produce heterochromatic siRNAs. They are the specificity determinants that guide the process of RNA-directed DNA methylation (RdDM), likely via the targeting of nascent long non-coding RNAs produced by a specialized DNA-dependent RNA polymerase, Pol V (reviewed by [35]).

Less attention has been paid to systematic annotation of individual heterochromatic siRNA loci, and there is no miRBase-type registry or database for these types of genes. Several groups have, however, reported the results of in-house computational approaches that defined heterochromatic siRNA loci on the basis of simple clustering methods coupled with analysis of heterochromatic siRNA mutants [36-38]. Two recent studies have described the role of the SHH1/DTF1 DNA-binding protein in guiding the formation of *Arabidopsis* heterochromatic siRNAs and in the process defined sets of heterochromatic siRNA loci. Law et al. [39*] defined ~12,500 heterochromatic siRNA loci by clustering of uniquely mapping 24 nt siRNAs. Similarly, Zhang et al. [40*] defined 4,187 loci comprised mainly of 24 nt RNAs that were strongly down-regulated in *dtf1* mutants. Both studies showed that *SHH1/DTF1* is a major regulator of heterochromatic siRNA levels. Importantly, SHH1/DTF1 is suggested to recruit Pol IV, which transcribes the precursors of heterochromatic siRNAs [35], to loci based upon the presence of H3K9 methylation marks [39*,40*]. These data suggest that prior deposition of repressive histone modifications is a pre-requisite for heterochromatic siRNA biogenesis.

Several lines of evidence indicate that heterochromatic siRNA gene annotation should not depend on a rigid siRNA size requirement of 24 nts. In cell suspension cultures [41] and pollen [42] *Arabidopsis* transposable elements that normally produce 24 nt heterochromatic siRNAs instead begin to produce appreciable amounts of 21-22 nt siRNAs. Nuthikattu et al. [43**] demonstrated that, upon global erasure of DNA methylation in the *Arabidopsis ddm1* mutant, 15 families of transposable elements begin to produce very high amounts of 21-22 nt siRNAs. These are dependent upon *RDR6*, which had previously been associated with secondary and phased siRNAs, but not heterochromatic siRNAs. The *RDR6*-dependent 21-22 nt siRNAs were capable of directing RdDM, making them *bona fide* heterochromatic siRNAs [43**]. Similarly, Marí-Ordóñez et al. [44**] also demonstrated that an epigenetically re-activated transposon, *EVD*, initially is targeted by 21-22 nt siRNAs. Over multiple generations of inbreeding, *EVD* is eventually silenced by RdDM. Interestingly, over the course of several generations, *EVD*-derived siRNAs transitioned from *RDR6*-dependent 21-22 nt siRNAs to Pol IV-dependent 24 nt siRNAs. Together, these studies suggest a model in which active transposable elements are first targeted by the secondary siRNA pathway, which makes 21-22 nt siRNAs that can cause both transcriptional and post-transcriptional silencing. Later, there is a gradual handoff to the 24 nt, Pol IV / Pol V heterochromatic siRNA pathway as the transcriptional silencing of the element becomes firmly entrenched. This implies that the prevalence of 24 nt heterochromatic siRNAs across many plant genomes represents a final ‘maintenance’ state for transposons and retroviruses that invaded long ago. There is evidence indicating that 21-22 nt ‘initiation’ state heterochromatic siRNA loci also exist in wild-type plants. Genome-wide analysis of DNA methylation in *Arabidopsis rdr6* mutants identified 138 loci with *RDR6*-dependent DNA methylation, most of which were associated with 21-22 nt siRNAs and distinct from the

DNA methylation caused by the canonical heterochromatic siRNA pathway [45]. In maize, which has a huge load of very active transposons, there are large numbers of 22 nt small RNAs which are not dependent on the canonical 24 nt heterochromatic siRNA pathway [46].

Resources for creating and disseminating annotations

A great number of programs geared specifically to *MIRNA* locus annotation exist, with several that are specialized for the unique features of plant *MIRNAs* [47-49]. Several related algorithms designed to detect the unique phasing signature of phased siRNA loci have also been described [28,29,50]. General purpose clustering methods that define loci of small RNA production based on small RNA-seq alignments also are available [51-54]. The UEA sRNA workbench [55] contains several stand-alone programs that individually address *MIRNA* annotation, general small RNA cluster identification, and phased siRNA locus annotation. Our program ShortStack [24*,56] generates annotations of *MIRNA* loci, other hpRNA loci, phased siRNA loci, and all other types of small RNA loci. Recent versions of ShortStack have added the capability to handle read-trimming and alignment of data [56], making it an integrated solution to generate small RNA gene annotations from raw small RNA-seq data.

Several web-based resources exist to disseminate plant small RNA gene annotations and related small RNA-seq alignment data (Table 1). As discussed above, miRBase [1] is the central repository and arbitrator for *MIRNA* loci from all species. The Meyers Lab maintains one of the most extensive small RNA web servers, primarily focused on plant species [57]. At present, 15 plant species are represented, each with easily queried databases of aligned small RNA-seq data, and custom-built genome browsers. Other web servers focus on small RNA-seq alignments and annotations for specified species [58-59]. To the best of our knowledge, the current web servers are primarily focused on providing and visualizing small RNA-seq alignment data, as opposed to the curation and dissemination of stable reference annotations (with the exception of *MIRNAs* from miRBase). To address this, we are developing a web server (plantsmallrnagenes.psu.edu) whose focus goes beyond delivery and visualization of alignment data by adding comprehensive reference annotations for small RNA-producing loci. As of this writing, the site hosts annotations for just two species (*Amborella trichopoda* and *Physcomitrella patens*), but large expansion is planned over the next year.

Conclusion

Enormous amounts of small RNA-seq data are now available for many plant species, and the barriers to obtaining even more data grow lower and lower. Much progress has been made in annotations, but this progress has been unevenly distributed, with *MIRNA* loci in particular receiving a disproportionate share of the attention. We believe that further efforts at comprehensive and consistent reference annotations of all types of small RNA producing genes, and improvements in the dissemination of such annotations, will greatly enhance the future of plant genomics. In particular, we look forward to the day when researchers seeking

to study small RNAs will be liberated from the need to “reinvent the wheel” with each analysis by generating their own *de novo* annotations of small RNA-producing genes.

Acknowledgments

Research in the Axtell Lab is currently supported by grants from the NSF (award 1121438) and the NIH (award R01 GM084051). We apologize to colleagues whose papers were not cited due to space restrictions or our ignorance.

References Cited

papers of special interest (*)

papers of outstanding interest (**)

1. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 2011; 39:D152–157. [PubMed: 21037258]
2. Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen X, Green PJ, et al. Criteria for annotation of plant MicroRNAs. *Plant Cell.* 2008; 20:3186–3190. [PubMed: 19074682]
3. Liu C, Axtell MJ, Fedoroff NV. The helicase and RNaseIIIa domains of Arabidopsis Dicer-Like1 modulate catalytic parameters during microRNA biogenesis. *Plant Physiol.* 2012; 159:748–758. [PubMed: 22474216]
4. Carbonell A, Fahlgren N, Garcia-Ruiz H, Gilbert KB, Montgomery TA, Nguyen T, Cuperus JT, Carrington JC. Functional analysis of three Arabidopsis ARGONAUTES using slicer-defective mutants. *Plant Cell.* 2012; 24:3613–3629. [PubMed: 23023169]
5. Manavella PA, Koenig D, Weigel D. Plant secondary siRNA production determined by microRNA-duplex structure. *Proc Natl Acad Sci U S A.* 2012; 109:2461–2466. [PubMed: 22308502]
6. Zhang X, Zhao H, Gao S, Wang WC, Katiyar-Agarwal S, Huang HD, Raikhel N, Jin H. Arabidopsis Argonaute 2 regulates innate immunity via miRNA393(*)-mediated silencing of a Golgi-localized SNARE gene, MEMB12. *Mol Cell.* 2011; 42:356–366. [PubMed: 21549312]
7. Manavella PA, Koenig D, Rubio-Somoza I, Burbano HA, Becker C, Weigel D. Tissue-specific silencing of Arabidopsis SU(VAR)3-9 HOMOLOG8 by miR171a. *Plant Physiol.* 2013; 161:805–812. [PubMed: 23204429]
8. Vaucheret H. AGO1 homeostasis involves differential production of 21-nt and 22-nt miR168 species by MIR168a and MIR168b. *PLoS ONE.* 2009; 4:e6442. [PubMed: 19649244]
- 9*. Jeong DH, Thatcher SR, Brown RSH, Zhai J, Park S, Rymarquis LA, Meyers BC, Green PJ. Comprehensive Investigation of MicroRNAs Enhanced by Analysis of Sequence Variants, Expression Patterns, ARGONAUTE Loading, and Target Cleavage. *Plant Physiology.* 2013; 162:1225–1245. This article demonstrates that heterogeneity in MIRNA processing is widespread in Arabidopsis, and that different miRNA isoforms can have differential accumulation according to tissue- or environment-specific cues. [PubMed: 23709668]
10. Yu B, Yang Z, Li J, Minakhina S, Yang M, Padgett RW, Steward R, Chen X. Methylation as a crucial step in plant microRNA biogenesis. *Science.* 2005; 307:932–935. [PubMed: 15705854]
- 11*. Zhai J, Zhao Y, Simon SA, Huang S, Petsch K, Arikait S, Pillay M, Ji L, Xie M, Cao X, et al. Plant MicroRNAs Display Differential 3′ Truncation and Tailing Modifications That Are ARGONAUTE1 Dependent and Conserved Across Species. *The Plant Cell.* 2013; 25:2417–2428. This article demonstrates the 3′ truncation and oligo-U tailing is common in hen1 mutants from several species, that 3′ truncation and oligo-U tailing occur for some miRNAs in the wild-type, and that these events occur on AGO-loaded miRNAs. [PubMed: 23839787]
12. Axtell MJ, Snyder JA, Bartel DP. Common functions for diverse small RNAs of land plants. *The Plant Cell.* 2008; 19:1750–1769. [PubMed: 17601824]
- 13*. Zhai J, Jeong DH, De Paoli E, Park S, Rosen BD, Li Y, González AJ, Yan Z, Kitto SL, Grusak MA, et al. MicroRNAs as master regulators of the plant NB-LRR defense gene family via the

production of phased, trans-acting siRNAs. *Genes Dev.* 2011; 25:2540–2553. This study demonstrates a large abundance of secondary, phased siRNAs arise from NB-LRR mRNAs in *Medicago* initiated by miR2118 and several other miRNA families. [PubMed: 22156213]

- 14**. Shivaprasad PV, Chen HM, Patel K, Bond DM, Santos BACM, Baulcombe DC. A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. *Plant Cell.* 2012; 24:859–874. This study demonstrates that members of the miR482/2118 superfamily initiate large amounts of phased, secondary siRNA accumulation from tomato NB-LRR genes. Additionally, pathogen infections by both a virus and a fungus are shown to correlate with decreased miR482 accumulation and increased NB-LRR mRNA accumulation. [PubMed: 22408077]
- 15*. Li F, Pignatta D, Bendix C, Brunkard JO, Cohn MM, Tung J, Sun H, Kumar P, Baker B. MicroRNA regulation of plant innate immune receptors. *Proc Natl Acad Sci U S A.* 2012; 109:1790–1795. This study demonstrates that tobacco miR6019 and miR6020 target the NB-LRR mRNA for the N gene, which causes accumulation of secondary, phased siRNAs. [PubMed: 22307647]
16. Johnson C, Kasprzewska A, Tennessen K, Fernandes J, Nan GL, Walbot V, Sundaresan V, Vance V, Bowman LH. Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome Res.* 2009; 19:1429–1440. [PubMed: 19584097]
- 17**. Xia R, Meyers BC, Liu Z, Beers EP, Ye S, Liu Z. MicroRNA Superfamilies Descended from miR390 and Their Roles in Secondary Small Interfering RNA Biogenesis in Eudicots. *The Plant Cell.* 2013; 25:1555–1572. This study demonstrates that several miRNA families considered to be unrelated by the standard miRNA annotation criteria in fact are descended from a common progenitor. This implies that extensive miRNA diversification may have obscured evolutionary relationships, and calls into question prior estimates of the rates of MIRNA gene birth and death in plants. [PubMed: 23695981]
18. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. *Plant Cell.* 2012; 24:4333–4345. [PubMed: 23136377]
19. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009; 25:1105–1111. [PubMed: 19289445]
20. Chuang CF, Meyerowitz EM. Specific and heritable genetic interference by double-stranded RNA in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 2000; 97:4985–4990. [PubMed: 10781109]
21. Wesley SV, Helliwell CA, Smith NA, Wang M, Rouse DT, Liu Q, Gooding PS, Singh SP, Abbott D, Stoutjesdijk PA, et al. Construct design for efficient, effective and high-throughput gene silencing in plants. *The Plant Journal.* 2001; 27:581–590. [PubMed: 11576441]
22. Fusaro AF, Matthew L, Smith NA, Curtin SJ, Dedic-Hagan J, Ellacott GA, Watson JM, Wang MB, Brosnan C, Carroll BJ, et al. RNA interference-inducing hairpin RNAs in plants act through the viral defence pathway. *EMBO Rep.* 2006; 7:1168–1175. [PubMed: 17039251]
23. Henderson IR, Zhang X, Lu C, Johnson L, Meyers BC, Green PJ, Jacobsen SE. Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature Genetics.* 2006; 38:721–725. [PubMed: 16699516]
- 24*. Axtell MJ. ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA.* 2013; 19:740–751. This study describes software designed for annotation of all types of small RNA genes based on reference-aligned small RNA-seq data. It also suggests that there are a large number of hpRNA loci in plant genomes. [PubMed: 23610128]
25. Dunoyer P, Brosnan CA, Schott G, Wang Y, Jay F, Alioua A, Himber C, Voinnet O. An endogenous, systemic RNAi pathway in plants. *The EMBO Journal.* 2010; 29:1699–1712. [PubMed: 20414198]
26. Slotkin RK, Freeling M, Lisch D. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat Genet.* 2005; 37:641–644. [PubMed: 15908951]
27. Fei Q, Xia R, Meyers BC. Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks. *The Plant Cell.* 2013; 25:2400–2415. [PubMed: 23881411]
28. Chen HM, Li YH, Wu SH. Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in *Arabidopsis*. *Proc Natl Acad Sci U S A.* 2007; 104:3318–3323. [PubMed: 17360645]

29. Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, Givan SA, Kasschau KD, Carrington JC. Genome-Wide Analysis of the RNA-DEPENDENT RNA POLYMERASE6/ DICER-LIKE4 Pathway in Arabidopsis Reveals Dependency on miRNA- and tasiRNA-Directed Targeting. *The Plant Cell*. 2007; 19:926–942. [PubMed: 17400893]
30. Nogueira FTS, Madi S, Chitwood DH, Juarez MT, Timmermans MCP. Two small regulatory RNAs establish opposing fates of a developmental axis. *Genes & Development*. 2007; 21:750–755. [PubMed: 17403777]
31. Axtell MJ, Jan C, Rajagopalan R, Bartel DP. A two-hit trigger for siRNA biogenesis in plants. *Cell*. 2006; 127:565–577. [PubMed: 17081978]
32. Arif MA, Fattash I, Ma Z, Cho SH, Beike AK, Reski R, Axtell MJ, Frank W. DICER-LIKE3 activity in *Physcomitrella patens* DICER-LIKE4 mutants causes severe developmental dysfunction and sterility. *Mol Plant*. 2012; 5:1281–1294. [PubMed: 22511605]
- 33*. Cho SH, Coruh C, Axtell MJ. miR156 and miR390 regulate tasiRNA accumulation and developmental timing in *Physcomitrella patens*. *Plant Cell*. 2012; 24:4837–4849. This study demonstrates that miR156 and miR390 levels are integrated by competing for tasiRNA production at the TAS6a/TAS3a non-coding RNA in *Physcomitrella patens*. [PubMed: 23263766]
34. Kallman T, Chen J, Gyllenstrand N, Lagercrantz U. A Significant Fraction of 21-Nucleotide Small RNA Originates from Phased Degradation of Resistance Genes in Several Perennial Species. *Plant Physiology*. 2013; 162:741–754. [PubMed: 23580593]
35. Wierzbicki AT. The role of long non-coding RNA in transcriptional gene silencing. *Current Opinion in Plant Biology*. 2012; 15:517–522. [PubMed: 22960034]
36. Mosher RA, Schwach F, Studholme D, Baulcombe DC. PolIVb influences RNA-directed RNA methylation independently of its role in siRNA biogenesis. *Proc Natl Acad Sci USA*. 2008; 105:3145–3150. [PubMed: 18287047]
37. Cho SH, Addo-Quaye C, Coruh C, Arif MA, Ma Z, Frank W, Axtell MJ. *Physcomitrella patens* DCL3 Is Required for 22–24 nt siRNA Accumulation, Suppression of Retrotransposon-Derived Transcripts, and Normal Development. *PLoS Genet*. 2008; 4:e1000314. [PubMed: 19096705]
38. Lee TF, Gurazada SGR, Zhai J, Li S, Simon SA, Matzke MA, Chen X, Meyers BC. RNA polymerase V-dependent small RNAs in Arabidopsis originate from small, intergenic loci including most SINE repeats. *Epigenetics*. 2012; 7:781–795. [PubMed: 22647529]
- 39*. Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AMS, Strahl BD, Patel DJ, Jacobsen SE. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature*. 2013; 498:385–389. Along with [40], this study provides evidence that the SHH1 DNA-binding protein is specific for histones H3K9 methylation mark and guides the positioning of RNA Pol IV, which in turn manufactures the precursors for 24 nt heterochromatic siRNAs. [PubMed: 23636332]
- 40*. Zhang H, Ma ZY, Zeng L, Tanaka K, Zhang CJ, Ma J, Bai G, Wang P, Zhang SW, Liu ZW, et al. DTF1 is a core component of RNA-directed DNA methylation and may assist in the recruitment of Pol IV. *Proceedings of the National Academy of Sciences*. 2013; 110:8290–8295. Along with [39], this study provides evidence that the SHH1 DNA-binding protein is specific for histones H3K9 methylation mark and guides the positioning of RNA Pol IV, which in turn manufactures the precursors for 24 nt heterochromatic siRNAs.
41. Tanurdzic M, Vaughn MW, Jiang H, Lee TJ, Slotkin RK, Sosinski B, Thompson WF, Doerge RW, Martienssen RA. Epigenomic consequences of immortalized plant cell suspension culture. *PLoS Biol*. 2008; 6:2880–2895. [PubMed: 19071958]
42. Slotkin RK, Vaughn M, Borges F, Tanurdzi M, Becker JD, Feijó JA, Martienssen RA. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*. 2009; 136:461–472. [PubMed: 19203581]
- 43*. Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, Slotkin RK. The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs. *Plant Physiology*. 2013; 162:116–131. This study demonstrates that, upon epigenetic activation, expressed transposons trigger production of RDR6-dependent 21-22 nt siRNAs, which are capable of directing RNA-directed DNA methylation. [PubMed: 23542151]
- 44*. Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O. Reconstructing de novo silencing of an active plant retrotransposon. *Nature Genetics*. 2013; 45:1029–1039. This

study examines a multi-generational time-course following the gradual establishment of silencing for an active retrotransposon. In the early phases, silencing is largely post-transcriptional and controlled by RDR6-dependent 21-22 nt siRNAs. In later generations, silencing is established at the transcriptional level, and is associated with 24 nt siRNAs. [PubMed: 23852169]

45. Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell*. 2013; 152:352–364. [PubMed: 23313553]
46. Nobuta K, Lu C, Shrivastava R, Pillay M, De Paoli E, Accerbi M, Arteaga-Vazquez M, Sidorenko L, Jeong DH, Yen Y, et al. Distinct size distribution of endogenous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. *Proc Natl Acad Sci U S A*. 2008; 105:14958–14963. [PubMed: 18815367]
47. Yang X, Li L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*. 2011; 27:2614–2615. [PubMed: 21775303]
48. Xie F, Xiao P, Chen D, Xu L, Zhang B. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Mol Biol*. 2012; 80:75–84.
49. Qian K, Auvinen E, Greco D, Auvinen P. miRSeqNovel: An R based workflow for analyzing miRNA sequencing data. *Molecular and Cellular Probes*. 2012; 26:208–211. [PubMed: 22609373]
50. De Paoli E, Dorantes-Acosta A, Zhai J, Accerbi M, Jeong DH, Park S, Meyers BC, Jorgensen RA, Green PJ. Distinct extremely abundant siRNAs associated with cosuppression in petunia. *RNA*. 2009; 15:1965–1970. [PubMed: 19776157]
51. MacLean D, Moulton V, Studholme DJ. Finding sRNA generative locales from high-throughput sequencing data with NiBLS. *BMC Bioinformatics*. 2010; 11:93. [PubMed: 20167070]
52. Pantano L, Estivill X, Marti E. A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. *Bioinformatics*. 2011; 27:3202–3203. [PubMed: 21976421]
53. Hardcastle TJ, Kelly KA, Baulcombe DC. Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinformatics*. 2012; 28:457–463. [PubMed: 22171331]
54. Chen CJ, Servant N, Toedling J, Sarazin A, Marchais A, Duvernois-Berthet E, Cognat V, Colot V, Voinnet O, Heard E, et al. ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics*. 2012; 28:3147–3149. [PubMed: 23044543]
55. Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, Schwach F, Dalmay T, Moulton V. The UEA sRNA Workbench: A Suite of Tools for Analysing and Visualising Next Generation Sequencing microRNA and Small RNA Datasets. *Bioinformatics*. 2012; 28:2059–2061. [PubMed: 22628521]
56. Shahid S, Axtell MJ. Identification and annotation of small RNA genes using ShortStack. *Methods*. 10.1016/j.ymeth.2013.10.004
57. Nakano M. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Research*. 2006; 34:D731–D735. [PubMed: 16381968]
58. Backman TWH, Sullivan CM, Cumbie JS, Miller ZA, Chapman EJ, Fahlgren N, Givan SA, Carrington JC, Kasschau KD. Update of ASRP: the Arabidopsis Small RNA Project database. *Nucleic Acids Research*. 2007; 36:D982–D985. [PubMed: 17999994]
59. Johnson C, Bowman L, Adai AT, Vance V, Sundaresan V. CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Research*. 2007; 35:D829–D833. [PubMed: 17169981]
60. Axtell MJ. Classification and comparison of small RNAs from plants. *Annual Review of Plant Biology*. 2013; 64:137–159.

Highlights

- Current microRNA annotation methods do not fully reflect emerging data
- Most plant small RNA genes are not microRNAs, and are unannotated
- Tools and websites for plant small RNA gene annotation are discussed

Figure 1.

Overview of plant small RNAs.

- A.** Processing and usage of plant small RNAs. Substrates are either the stems of hairpins (top left), or long double-stranded RNA (bottom left) which is often synthesized by an RNA-dependent RNA polymerase (RDR). Substrates are processed by Dicer-Like (DCL) proteins to yield initial duplexes. One of the two strands is bound to Argonaute (AGO) protein, and then guides the AGO protein to RNA targets based on complementarity.
- B.** Schematic of microRNA biogenesis from a hairpin precursor. *MIRNA* loci are defined as hairpins that produce a discrete initial duplex (the miRNA/miRNA* duplex). Red: miRNA, blue: miRNA*.

- C.** Schematic of small RNA biogenesis from a non-microRNA hairpin precursor, which is defined as small RNA-producing hairpin that does not meet the definition of a *MIRNA* locus.
- D.** Schematic of the biogenesis of secondary, phased siRNAs, which are defined by small RNA-directed cleavage of a long precursor, followed by dsRNA synthesis and dicing.
- E.** Schematic of heterochromatic siRNAs, which are defined by accumulation from intergenic regions and association with repressive DNA and histone modifications.

B-E: Modified from [60].

Figure 2.

MIRNA hairpins produce more than one product.

- A.** Schematic of a typical *MIRNA* locus with aligned reads from small RNA-seq, and explanation of terms and calculations.
- B.** Distribution of precision^{ann} values from *Arabidopsis MIRNA* loci with respect to miRBase 20. Based on genome alignment of a small RNA-seq dataset comprised of NCBI GEO GSM738731 and GSM738727.
- C.** As in B, except for precision^{max} values.
- D.** Frequency of concordance between miRBase 20 annotations of the mature miRNA, and the observed most abundant RNA for the small RNA-seq data.
- E.** As in B, except using small RNAs from an AGO1-IP experiment (NCBI GEO GSM989351), and restricting the analysis to *MIRNA* loci annotated with a mature miRNA 21 nt in length with a 5' U.
- F.** As in E, except for precision^{max} values..
- G.** Frequency of concordance as in D, except for AGO1-IP data and restricting the analysis to *MIRNA* loci annotated with a mature miRNA 21 nt in length with a 5' U.

Figure 3.

The annotation gap: comparison of observed expression data to annotations for small RNAs (NCBI GEO GSM738731 and GSM738727) and polyA+ RNAs (NCBI GEO GSM946222 and GSM946223) in *Arabidopsis*. See text for methodological details.

- A.** Area-proportional Venn diagram showing the extent (number of nts) of significant (defined as a coverage of ≥ 0.1 read per million) polyA+ RNA (RNA-seq) and small RNA-seq expression in the *Arabidopsis* genome.
- B.** Area-proportional Venn diagrams illustrating the overlap between areas of significant small RNA-seq or RNA-seq expression and annotated regions in *Arabidopsis* (left: small RNA-seq vs. miRBase, right: RNA-seq vs. TAIR10 genes including introns).

- C. Pie charts illustrating the proportion of aligned small RNA-seq reads overlapping *MIRNA* annotation (left), or the proportion of RNA-seq reads overlapping TAIR10 gene annotations including introns (right) for *Arabidopsis*.

Table 1

Selected websites that disseminate plant small RNA alignments and/or annotations.

Site Name	URL	Species currently present	Comments	Citation
miRBase	http://www.mirbase.org/	72 plants (as of version 20)	Disseminates <i>MIRNA</i> hairpin and mature miRNA annotations for all species, including plants.	[1]
University of Delaware SBS databases	http://mpss.udel.edu/	15 plant species	Small RNA-seq, RNA-seq, PARE/ degradome, and other high-throughput datasets with search functions and a custom-built genome browser for each species	[57]
ASRP	http://asrp.danforthcenter.org/	<i>Arabidopsis thaliana</i>	Disseminates small RNA-seq datasets and features a genome-browser.	[58]
CSRDB	http://sundarlab.ucdavis.edu/smrnas/	Maize and rice	Queryable small RNA-seq data along with target predictions and genome browsers	[59]
The plant small RNA genes web server at Penn State	http://plantsmallrnagenes.psu.edu/	<i>Physcomitrella patens</i> and <i>Amborella trichopoda</i>	Disseminates global reference annotations of small RNA producing genes (all types), along with full datasets and genome browsers.	This work