COMMENTARY

Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data

Michael J. Axtella and Blake C. Meyersb,c
aThe Pennsylvania State University, Department of Biology and Huck Institutes of the Life Sciences, University Park, Pennsylvania 16802
bDonald Danforth Plant Science Center, St. Louis, Missouri 63132
cUniversity of Missouri-Columbia, Division of Plant Sciences, Columbia, Missouri 65211

ORCID IDs: 0000-0001-8951-7361 (M.J.A.); 0000-0003-3436-6097 (B.C.M.)

MicroRNAs (miRNAs) are ~21-nucleotide-long regulatory RNAs that arise from endonucleolytic processing of hairpin precursors. Many function as essential posttranscriptional regulators of target mRNAs and long noncoding RNAs. Alongside miRNAs, plants also produce large numbers of short interfering RNAs (siRNAs), which are distinguished from miRNAs primarily by their biogenesis (typically processed from long double-stranded RNA instead of single-stranded hairpins) and functions (typically via roles in transcriptional regulation instead of posttranscriptional regulation). Next-generation DNA sequencing methods have yielded extensive data sets of plant small RNAs, resulting in many miRNA annotations. However, it has become clear that many miRNA annotations are questionable. The sheer number of endogenous siRNAs compared with miRNAs has been a major factor in the erroneous annotation of siRNAs as miRNAs. Here, we provide updated criteria for the confident annotation of plant miRNAs, suitable for the era of “big data” from DNA sequencing. The updated criteria emphasize replication and the minimization of false positives, and they require next-generation sequencing of small RNAs. We argue that improved annotation systems are needed for miRNAs and all other classes of plant small RNAs. Finally, to illustrate the complexities of miRNA and siRNA annotation, we review the evolution and functions of miRNAs and siRNAs in plants.
deposition of DNA methylation on the adjacent genomic DNA, ultimately leading to the reinforcement of heterochromatic histone marks.

Readers interested in the details of biogenesis and molecular functions of miRNAs, phasiRNAs, and hc-siRNAs should consult more comprehensive or specialized reviews (Rogers and Chen, 2013; Fei et al., 2013; Matzke and Mosher, 2014). Here, we focus on the new knowledge and challenges that have arisen from the application of high-throughput small RNA sequencing (sRNA-seq) to the discovery and study of these three major classes of plant small RNAs in diverse plants. We also envision this article as an update to a decade-old publication that laid out the criteria for the annotation of plant miRNAs (Meyers et al., 2008), taking into account the many technological and conceptual advances made since that time.

UPDATE ON CRITERIA FOR PLANT miRNA ANNOTATIONS: THE URGENT NEED TO MINIMIZE FALSE POSITIVES

The first coordinated effort to define community standards in miRNA annotations relied on a simple set of criteria involving a combination of evidence of expression (accumulation on an RNA gel blot or in a cDNA sequencing project) and biogenesis (presence of a hairpin, conservation of the miRNA, or reduction in accumulation in a dicer mutant background) (Ambros et al., 2003). This first effort occurred well before the full complexity of endogenous small RNAs was understood and was later found by many to be inadequate, especially in the context of sRNA-rich plants (for example, see Meyers et al., 2008). In particular, the accumulation of a small RNA on an RNA gel blot merely proves it accumulates but does not differentiate between miRNAs, endogenous siRNAs, and partially degraded fragments of RNA unrelated to any gene-regulatory mechanism. Conservation is not definitive as a criterion for annotation both because there is no a priori reason that a miRNA need be conserved (and, indeed, many appear not to be; see below) and because there are other types of small RNAs that are conserved. Also, reduction in a dicer mutant is not definitive in plants because plants express multiple DCL genes with partially redundant roles in both miRNA and siRNA biogenesis (Gascioli et al., 2005; Rajagopalan et al., 2006), as well as in other events such as tRNA processing (Cole et al., 2009; Martinez et al., 2017). In other words, reduction or loss in a dcl mutant proves the RNA is not just a random degradation product unrelated to gene regulation, but does not categorically differentiate between miRNAs and endogenous siRNAs.

Careless application of the 2003 criteria to modern deep sequencing data would have led to a deluge of false-positive miRNA annotations. Realizing this, in 2008, we coordinated the publication of a consensus opinion among many plant small RNA researchers on a better set of criteria for plant miRNA annotations (Meyers et al., 2008). The 2008 criteria essentially stated that the only necessary and sufficient criteria for annotating a miRNA locus is the demonstration of precise processing of a miRNA/miRNA* duplex from a computationally predicted hairpin RNA precursor that met certain minimal structural criteria (Table 1). One of the major requirements in terms of making a novel annotation is the sequencing of the exact miRNA*. miRNA* abundance may be one or two orders of magnitude less than that of the mature miRNA, so the requirement of miRNA* sequencing necessitates deep sequencing coverage.

In our opinion, the 2008 criteria for plant miRNA annotations have held up quite well in the last 10 years. However, our own research experiences in the intervening time suggest the following improvements (summarized in Table 1): (1) The hairpin of a miRNA precursor is a defining characteristic and a component of all reliable miRNA identification algorithms. Using conserved miRNAs as exemplars, there are key features of the foldback or hairpin that are important for Dicer recognition and consistent processing (Cuperus et al., 2011; Bologna et al., 2013a, 2013b; Chorostecki et al., 2017): (i) typically a single miRNA:miRNA* duplex (thus far not more than three duplexes, a number observed in miR319/159); (ii) the region of this foldback that gives rise to the miRNA duplex should not contain secondary stems or large internal loops (larger than five nucleotides) that interrupt the miRNA:miRNA* duplex; (iii) while miRNA foldbacks of plants are often longer than those of animals, most foldbacks are just a few hundred nucleotides. These structural characteristics govern processing including the base-to-loop versus loop-to-base direction of duplex production. Foldbacks longer than 300 nucleotides should not be annotated as miRNA-producing loci.

(2) Expression of one or more miRNA/miRNA* duplexes should be observed by high-throughput sRNA-seq, as opposed to reliance on RNA gel blotting. RNA gel blotting for small RNAs carries the risk of probe cross-hybridization. sRNA-seq methods are now routine and should be the standard method of expression analysis used to support novel miRNA annotations in plants.

(3) and (4) Changes in structural criteria: Up to five mismatched positions between the miRNA and miRNA*, only three of which can be asymmetrical bulged, should be allowed. These changes correspond to observed extremes among validated miRNAs. Note that these criteria were separate in the 2008 “rules” (Table 1) but are similar and thus combined in the 2018 criteria.

(5) Precision of miRNA/miRNA* processing and positional variants. The 2008 criteria suggested that 75% or more of all aligned small RNAs at a locus should correspond to the exact miRNA or miRNA* sequences. Our experiences with sRNA-seq data from many species since then indicate this is too strict. In particular, we observe that 5’ and especially 3’ positional variants of miRNAs and miRNA* are quite common in sRNA-seq data. These variants likely result from a combination of DCL cleavage at different positions (including DCL imprecision, perhaps a characteristic of recently evolved miRNAs; Bologna et al., 2013b), post-dicing modifications of the 3’ end (Zhai et al., 2013), and sample degradation. Thus, the revised calculation of precision includes the counts of the exact mature miRNA and its corresponding exact miRNA*, as well as all one-nucleotide variants thereof. A one-nucleotide variant is defined as a small RNA for which both the 5’ and 3’ ends are each within one


Table 1. Updated Criteria for Plant miRNA Annotations

<table>
<thead>
<tr>
<th></th>
<th>2008 Criteria</th>
<th>2018 Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>One or more miRNA/miRNA* duplexes with two-nucleotide 3’ overhangs</td>
<td>Add requirements that exclude secondary stems or large loops in the miRNA/miRNA* duplex and limit precursor length to 300 nucleotides</td>
</tr>
<tr>
<td>2</td>
<td>Confirmation of both the mature miRNA and its miRNA*</td>
<td>Disallow confirmation by blot; sRNA-seq only</td>
</tr>
<tr>
<td>3</td>
<td>miRNA/miRNA* duplex contains ≤4 mismatched bases</td>
<td>Up to five mismatched positions, only three of which are nucleotides in asymmetric bulges</td>
</tr>
<tr>
<td>4</td>
<td>The duplex has at most one asymmetric bulge containing at most two bulged nucleotides</td>
<td>Up to five mismatched positions, only three of which are nucleotides in asymmetric bulges</td>
</tr>
<tr>
<td>5</td>
<td>≥75% of reads from exact miRNA or miRNA*</td>
<td>Include one-nucleotide positional variants of miRNA and miRNA* when calculating precision</td>
</tr>
<tr>
<td>6</td>
<td>Replication suggested but not required</td>
<td>Required; novel annotations should meet all criteria in at least two sRNA-seq libraries (biological replicates)</td>
</tr>
<tr>
<td>7</td>
<td>Homologs, orthologs, and paralogs can be annotated without expression data, provided all criteria met for at least one locus in at least one species</td>
<td>Homology-based annotations should be noted as provisional, pending actual fulfillment of all criteria by sRNA-seq</td>
</tr>
<tr>
<td>8</td>
<td>miRNA length not an explicit consideration</td>
<td>No RNAs &lt;20 nucleotide or &gt;24 nucleotides should be annotated as miRNAs. Annotations of 23- or 24-nucleotide miRNAs require extremely strong evidence.</td>
</tr>
</tbody>
</table>

nucleotide of the corresponding ends of the exact miRNA or miRNA*. Thus, for each miRNA and miRNA*, there are eight possible one-nucleotide variants. Precision is thus redefined to be the sum of reads for the miRNA, miRNA*, and their respective one-nucleotide variants divided by the total number of reads aligned to the locus. When calculated in this manner, precisions for high-confidence microRNA loci tend to be around 90%; the minimum allowable percentage remains 75%.

(6) Replication: Annotations of novel miRNA families require support from independent small RNA-seq libraries. That is, all of the updated criteria must be fulfilled in at least two distinct sRNA-seq libraries (not technical replicates). Ideally, these replicates would come from different tissues, developmental stages, or treatments, as miRNAs show a higher degree of qualitative reproducibility across any of these than siRNAs. However, biological replicates of the same tissuestage/treatment could suffice for miRNAs for which expression is truly specific to one condition.

(7) Annotation by homology: We feel that any annotations based on homology alone to a known miRNA family should be regarded as provisional, and clearly marked so, until such time as all the required sRNA-seq based criteria are met for the newly annotated locus. This will slow down the rate of propagation of errors caused by homology-based annotations of families erroneously annotated in the first place.

(8) miRNA length: One effective method to minimize false positive annotations is to categorically disallow annotations of miRNAs that are not the expected size. We are not aware of any evidence demonstrating that miRNAs <20 or >24 nucleotides are loaded onto any plant AGO proteins. Unless such evidence emerges, RNAs <20 or >24 nucleotides in length should never be annotated as miRNAs. hc-siRNAs are mostly 23 or 24 nucleotides in length, and they are extremely numerous in many plant sRNA-seq libraries. In contrast, bona fide 23- or 24-nucleotide miRNAs are very rare. The huge number of 23- or 24-nucleotide hc-siRNAs coupled with scant experimental evidence for 23- or 24-nucleotide miRNAs suggests that avoiding annotations of 23- or 24-nucleotide RNAs as miRNAs is prudent. Any exceptions should have truly extraordinary evidence of miRNA biogenesis and function, including precise miRNA/miRNA* accumulation in four or more separate sRNA-seq libraries and ideally a direct demonstration that they posttranscriptionally regulate target RNAs in a miRNA-like manner (i.e., translational repression, target cleavage, and/or phasiRNA accumulation). Good examples of bona fide 24-nucleotide miRNAs that fulfill these enhanced criteria can be found in rice (Oryza sativa; Wu et al., 2010). One potential caveat to this rule is that it remains possible that some plant species generate numerous 23- or 24-nucleotide miRNAs, but these species have not yet been characterized or had extensive small RNA analyses.

Figure 1 shows an example of two Arabidopsis loci analyzed using sRNA-seq data from inflorescences. While the locus ath-MIR399b meets all of the revised criteria described here, ath-MIR405a does not.

Why is avoiding false-positive miRNA annotations of critical importance? The major issue is that, given the complexity of plant endogenous small RNA populations, even very small rates of false-positive miRNA annotations lead to overwhelming numbers of bad annotations. Many computational tools for plant miRNA discovery have been described, but they vary widely in performance. Lei and Sun (2014) performed an informative benchmarking experiment in which the same small RNA data sets from Arabidopsis were used as input to seven different miRNA annotation tools (Figure 2). In these analyses, we assume for the sake of argument that the miRNA loci found by a program that did not already exist in miRBase (the central miRNA registry; see below) are false positives. Strikingly, several tools found more than 1000 false
positives during their analyses, and consequently have poor precision and accuracy. Lei and Sun’s analysis suggested that miR-PREFeR (Lei and Sun, 2014) and ShortStack (Johnson et al., 2016) performed the best. However, even the best tools had false positives, along with accuracies that rarely surpassed 60% (Figure 2). Plus, the software landscape is continually evolving so it is likely that these tools will improve or better ones may yet emerge. In light of progress in machine learning approaches, a possibility in the future of miRNA annotation is to efficiently sort small RNAs and classify miRNAs with a higher

**Figure 1.** Examples of Valid and Invalid MiRNA Loci.

Small RNA-seq data were from GSE105262 (Polydore and Axtell, 2017). Left: Arabidopsis (ath) MIR399b, a locus judged valid under the new criteria. Right: ath-MIR405a, a locus judged invalid under the new criteria.
degree of accuracy than current methods, perhaps even without a reference genome. Indeed, the recent description of the mir-\textit{novo} tool (Vitsios et al., 2017) is promising, although its performance on plant data sets is diminished relative to animals. Whatever tool or internal procedure is used to annotate miRNAs from sRNA-seq data, users should exercise caution and common sense, and ask whether the procedure conforms to all of the annotation guidelines described above. In particular, any analysis that results in thousands of new miRNA annotations is quite likely to be comprised almost exclusively of false positives (Taylor et al., 2017), because even in intensively studied species like Arabidopsis, at most a few hundred miRNAs are known. Researchers, peer reviewers, and consumers of the published literature should be skeptical of a report...
that describes 100+ “new” miRNA loci in a given plant species.

THE STATE OF miRBase

miRBase (www.miRBase.org) is the centralized registry and database for miRNA annotations across all studied species (Kozomara and Griffiths-Jones, 2014). Its primary purposes are to assign consistent nomenclature to miRNA loci and to allow easy community access to all known miRNAs and their sequences. As of this writing, miRBase is at version 21, containing 6942 miRNA hairpin annotations from 72 different land plant species; unfortunately, it has been over 3 years since this last release. miRNAs are placed into families based on the similarity of the mature miRNA sequences; individual families are assigned distinct numbers in the order of their discovery. Multiple loci from the same family can be present in a single genome, and families are conserved among plant species. As a result, the 6942 hairpins represent 2408 distinct miRNA families of land plants.

It is important to realize that miRBase is not the primary gatekeeper in terms of enforcing the quality of miRNA annotations. Instead, miRBase essentially collates and standardizes annotations from the peer-reviewed literature. The burden of quality control for miRNA annotations instead falls on researchers, peer reviewers, and editors. Because of the reliance on peer review and its inherently inconsistent results, and because of the large risk of false positive annotations (see above), the overall quality of miRBase is suboptimal. Taylor et al. (2014) provided an important retrospective analysis of all of the land plant miRNA annotations present in miRBase version 20. Their analysis, based on processing patterns inferred from sRNA-seq data, found that 1993 out of the 6172 (32%) land plant miRNA loci lacked convincing supporting evidence. These dubious loci were disproportionately “singleton” families; that is, miRNA families represented by just a single locus in a single species. Thus, in their analysis, Taylor et al. (2014) marked as questionable 1351 of the 1802 (75%) of the land plant miRNA families in miRBase version 20. To the uninitiated, the idea that three out of four plant miRNA families are false annotations is probably a shock. Nonetheless, the criteria that Taylor et al. (2014) used are largely consistent with the best-practice methods we outline above (Table 1); thus, these estimates of the problems with miRBase annotations are likely close to the mark. miRNA annotations in miRBase for human and animal genomes are also quite problematic for many of the same reasons (Fromm et al., 2015).

The curators of miRBase have recognized the growing issue of false positive miRNA annotations (Kozomara and Griffiths-Jones, 2014), and they have taken action. Beginning with the release of version 20, miRBase allows crowd-sourced comments on individual annotations, including the ability to “upvote,” “downvote,” and comment on annotations, and in some cases links to Wikipedia entries describing each locus. miRBase has also designated a subset of miRNA annotations as “high-confidence.” High-confidence loci are designated based on the pattern of read alignments from reference-aligned sRNA-seq data. The criteria are similar in spirit to our recommended best practices (compare Tables 1 and 2). Strikingly, there aren’t that many high-confidence land plant miRNAs: Just 587 of the 6942 (8.5%) of the land plant miRNA loci in miRBase 21 are marked as high confidence. If we designate an miRNA family as high confidence provided it has at least one high-confidence locus, we observe that just 227 of the 2408 (9.4%) land plant miRNA families are “high confidence” as of miRBase 21. Clearly, available data support a conclusion that a great many miRBase entries, perhaps the majority, are questionable. Mere presence of an annotation in miRBase absolutely should not be taken to prove that a miRNA annotation is “real”, and instead, miRNA annotations must withstand the test of time and secondary analysis by independent groups and data sets. Importantly, miRBase also displays aligned small RNA-seq data for several species. When available, inspection of these data can help users decide for themselves the reliability of miRNA annotations.

Other retrospective efforts to reevaluate miRBase miRNA annotations have focused on individual species, including rice (Jeong et al., 2011) and soybean (Glycine max; Arikit et al., 2014). In these cases, extensive manual curation coupled with computational analyses were used to assess prior annotations in light of deep sRNA-seq coverage. These efforts are valuable, but have limits. One of the issues is that miRBase is not designed for bulk removal of clearly erroneous annotations (although the idea of separating high-confidence annotations from the general population is promising). Another issue with the static-list approach is that the effort required is large, and the fast-moving pace of data accumulation can change the results quickly. New sRNA-seq data arrive in public archives almost

---

### Table 2. miRBase Criteria for High-Confidence miRNA Annotations (Reproduced from Kozomara and Griffiths-Jones, 2014)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>At least 10 sRNA-seq reads aligned with no mismatches to both the mature miRNA and the miRNA*</td>
</tr>
<tr>
<td>2</td>
<td>The most abundant reads from each arm of the hairpin precursor must pair in the miRNA/miRNA* duplex with a 0- to 4-nucleotide overhang at their 3’ ends</td>
</tr>
<tr>
<td>3</td>
<td>At least 50% of reads mapping to each arm of the hairpin precursor must have the same 5’ end</td>
</tr>
<tr>
<td>4</td>
<td>The predicted hairpin structure must have a folding free energy of &lt;0.2 kcal/mol/nucleotide</td>
</tr>
<tr>
<td>5</td>
<td>At least 60% of the bases in the miRNA and miRNA* must be paired in the predicted hairpin structure</td>
</tr>
</tbody>
</table>
daily. Thus, over time, initially questionable miRNA annotations might become better supported, while others that initially looked strong might be questioned. Given the volume of incoming sRNA-seq data, it is unreasonable to expect any one person or group (like the miRBase curators) to maintain a fully up-to-date database of all available aligned reads across all plant species, as well as making available what would be a constantly evolving set of evaluations.

Finally, a problem of growing concern is the large length of time between updates to miRBase. As of this writing (the last week of 2017), no updates to miRBase have been released in more than 3 years, although a social media announcement has promised a new update soon. During this time, many new miRNA submissions have been submitted to miRBase and received official registry numbers, but none have yet been released by the database. This presents a serious problem in that it has become difficult to assert with confidence that “new” miRNA annotations are truly new; it is not feasible to scour every single published paper on miRNAs and hunt through supplemental data tables to find all previously annotated miRNA sequences for all species. Perhaps part of the issue in maintaining miRBase is simply the volume of submissions and the workload required for manual curation of each newly-submitted list of miRNAs. Slow release of new annotations may also have the effect of discouraging submission of new sequences.

We suggest that either miRBase or the community should consider alternatives to the current system. We advocate for the development of a fully automated system that would permit uploads of candidates in a standardized format without requiring direct curator input. Quality control would be maintained by keeping the requirement that new miRNA annotations are part of an accepted or published peer-reviewed article and by requiring that submitters’ names and contact information be publicly listed next to their annotations. The automated system could also be designed to enforce the modified requirements we outline above. In particular, it could require the underlying sRNA-seq alignments to be uploaded, followed by automated analysis of the prospective miRNA annotations. Users could receive a report on each locus, and with loci accepted only if they pass the analysis. Accepted annotations would, at the submitter’s discretion, become instantly available or available upon publication of the associated study, with names automatically generated and registered. We also envision a system in which other users could evaluate and rank previous annotations based on the criteria above (e.g., Table 1). Users could influence rankings of individual miRNAs via submission of additional data or reports.

CONSERVED LAND PLANT miRNAs

Understanding the challenges inherent in miRNA annotation requires some background on their patterns of conservation across plant genomes. Floyd and Bowman (2004) made the seminal observation that the predicted miRNA target sites for miR166 (one of the first miRNAs discovered in Arabidopsis) were conserved in all major land plant lineages, including angiosperms, gymnosperms, ferns, lycopsods, mosses, liverworts, and hornworts. A subsequent microarray study showed that several miRNA families first described in Arabidopsis were detectable in ferns, gymnosperms, magnoloids, and monocots, with a few also detectable in a lycopod and a moss (Axtell and Bartel, 2005). The subsequent 10+ years have seen numerous studies that have applied sRNA-seq to the task of identifying miRNAs in diverse plant lineages. The key enabling technologies in these studies have been the development of highly parallel next-generation sequencing methods and the assembly of complete genome sequences for several key plant species. From an evolutionary perspective, the genome sequences of Amborella trichopoda (an angiosperm that is sister to all other extant angiosperms; Amborella Genome Project, 2013), the lycopod Selaginella moellendorffii (Banks et al., 2011), and of the moss Physcomitrella patens (Rensing et al., 2008) have proven especially critical because they represent key clades in land plant evolution that are generally undersampled in terms of genomic resources.

Defining a “conserved” miRNA is an inherently slippery concept. When applied to miRNAs, conservation generally refers to the accumulation of identical or near-identical (one or two nucleotide differences) mature miRNA sequences, although the miRNA* and some flanking hairpin sequence also sometimes demonstrate conservation (Chorostecki et al., 2017). One of the key issues is at what taxonomic level conservation is observed. For the sake of this discussion, we define a “conserved” land plant miRNA family as one that has been annotated in miRBase in at least two of the eight following major taxonomic divisions shown in Figure 3: eudicots-rosids, eudicots-asterids, eudicots-ranunculids, monocots (mostly Poaceae), Amborella, gymnosperms, lycophytes, and bryophytes. We note that the core eudicots (rosids and asterids) and monocots dominate miRBase. By contrast, ranunculid eudicots, lycophytes, and bryophytes are each represented by just a single species (Aquilegia caerulea, S. moellendorffii, and P. patens, respectively). At present, miRBase lacks representatives of several divisions, such as magnoloids, monilophytes (ferns and fern allies), hornworts, and liverworts. We further constrain the definition to minimize false positives by requiring that at least one of the annotations is designated “high-confidence” in miRBase 21. Using these criteria, we see that 36 miRNA families are “conserved” (Figure 3). These include nine families (miR156, miR160, miR166, miR171, miR319, miR390, miR477, miR529, and miR535) with high-confidence annotations in both P. patens and at least one angiosperm. These nine miRNAs most likely arose prior to the existence of the last common ancestor of all land plants and have been conserved in most or all diversified lineages.

The uneven sampling density across these divisions means that apparent patterns of loss in a lineage should be interpreted skepticismally; they may well be artifacts of undersampling. This is especially true in the ranunculid eudicots (and other sister taxa to the “core” eudicots), lycophytes, and gymnosperms. Indeed, a study of spruce small RNAs (Xia et al., 2015a) expanded the roster of conserved miRNA families in gymnosperms but is not yet reflected in the current version of miRBase.
Figure 3. Conserved miRNA Families in Land Plants Based on miRBase 21 Annotations.

Annotations were binned into one of eight major taxonomic groupings. Top bar charts illustrate counts of the number of species, number of miRNA families, and number of high-confidence miRNA families in each group. The central heat map shows annotations for “conserved” miRNA families, which are defined here as those annotated in at least two of the eight groups and which have at least one high-confidence annotation. The bottom cladogram illustrates the approximate divergence times (based on Magallón et al. [2013] and Wang et al. [2009]) of the eight groups. MYBP, million years before present.
nor is a more recent study reporting extensive small RNA-seq data from many lycophyte and fern species (You et al., 2017). Therefore, there are still likely to be widely conserved miRNAs that are not yet recognized as such, due to as yet incomplete sampling of plant species and their small RNAs.

Proper miRNA annotations require both sRNA-seq data and the availability of a reference genome. This is because precursor hairpin structure is a key required feature needed for annotation. This restricts de novo annotations of new miRNA families to species with sequenced genomes. Nonetheless, survey sRNA-seq data from species that lack a sequenced genome can still be useful to illuminate patterns of miRNA evolution and conservation. The disadvantages with this approach are that only those miRNA families that can be confidently annotated from “anchor” species that do have reference genomes are countable. Chávez Montes et al. (2014) reported a survey of land plant miRNA evolution based on an extensive sRNA-seq effort from 31 diverse vascular plant species, including one fern and three gymnosperms. Several patterns not apparent from miRBase annotations alone were identified. Perhaps most importantly, the observation of highly abundant miRNA reads from multiple species can buttress the annotation confidence for clades that are lightly sampled in miRBase. Specifically, the observation of high miRNA read counts from several miRNA families in multiple gymnosperm and diverse angiosperm species (Chávez Montes et al., 2014) gives much more confidence in the existing miRBase annotations, which have few to no high-confidence annotations in miRBase at present (Figure 3). A large-scale, reference-genome-free, sRNA-seq effort from lycophytes and ferns was similarly able to buttress confidence of many more miRNA annotations (You et al., 2017).

MECHANISMS AND RATES OF PLANT miRNA EMERGENCE

As we demonstrate above and from numerous studies of miRNAs in divergent lineages, it is clear that individual miRNAs have emerged steadily and in parallel to the evolution of land plants. Publications describing the mechanisms by which plant miRNAs emerge are numerous, so rather than covering the topic in detail, we will simply note the consensus points. In miRBase version 21, there are 2026 land plant miRNA families (out of 2408 total families) that are only annotated in a single species; of those, just 106 families are “high-confidence”; in other words, there are a variety of lineage-specific miRNAs plus a subset of miRNAs conserved to varying degrees. As we describe above, many apparently “species-specific” miRNAs may be misannotated sRNAs; ignoring those, there are still many nonconserved, lineage-specific miRNAs. Where do these come from and what is the process by which they emerge? Two primary mechanisms are clear: (1) spontaneous genomic formation via duplication of inverted repeats, and (2) gradual evolutionary shifts in miRNA sequences, in parallel with target gene divergence. In the first case, homology to target genes in miRNA precursors outside of the miRNA:miRNA* duplex supports inverted repeat formation as a step to miRNA generation (Allen et al., 2004); this is readily apparent among the highly redundant families that target nucleotide binding site, leucine-rich repeat (NLR) resistance genes (Zhang et al., 2016). The second case requires identification of sequence conservation across precursors of miRNAs, some of which may break the rules we describe above for designation of families. One example is the miR7122 “super-family” that includes miR173, miR7122, and miR1509 (among others) and shares a core sequence of just 13 nucleotides with the much more ancient miRNA, miR390 (Xia et al., 2013). The miR4376 super-family was a “stepping stone” to the identification of the emergence of miR7122, in that the miR7122-miR390 homology was weak and cryptic without knowing the intermediate homology of miR4376. The key lesson is that some miRNAs have “hidden conservation,” as yet unknown due to sampling bias; as new species’ genomes are examined that fill in the gaps between the limited number of sequenced plant genomes and characterized miRNAs, connections between “new” and “conserved” miRNAs are likely to be found. In the near term, this confounds efforts to estimate birth-and-death rates of plant miRNAs. Overall, while it is clear that many apparently lineage-specific miRNAs are likely erroneous annotations (Taylor et al., 2014), it also remains clear that many others are real. miRNAs without obvious conservation will continue to be a major part of future annotation efforts in plants.

CONSERVATION AND IDENTIFICATION OF miRNA TARGETS

For the most widely conserved plant miRNAs, their targets are also conserved (for a good list of these well-known conserved targets, see Table 1 in Jones-Rhoades, 2012). Indeed, purifying selection offers an easy explanation for conservation of miRNA sequences and their target sites: If a given miRNA-target interaction is critical for fitness, most mutations would be deleterious, as they would be most likely to decrease the complementarity between the miRNA and the target. This is especially true for miRNAs that have multiple important targets, as simultaneous mutations in the miRNA and all of the target sites that maintain base-pairing are highly unlikely.

Layered on top of this static continuity of old miRNAs and their old target sites is the observation that old miRNAs can pick up new, lineage-specific targets. miR396, which appears to have arisen before the last common ancestor of the seed plants (Figure 3) provides some striking examples. The “old” miR396 targets are Growth Regulating Factor (GRF) miRNAs, which have been confirmed as targets in the eudicot Arabidopsis (Jones-Rhoades and Bartel, 2004), the monocot rice (Li et al., 2010), and predicted as targets in the lyco-pod S. moellendorffii (Debernardi et al., 2012). In the Brassicaceae and Cleomaceae (both in the Rosids supergroup of eudicots), miR396 also targets Basic Helix-Loop-Helix 74 (bHLH74) homologs (Debernardi et al., 2012), while in sunflower (Helianthus annuus); in the asterid supergroup of eudicots) it has gained a WRKY transcription factor target (Giacomelli et al., 2012). Gains of lineage-specific targets have also been described for the ancient miR156, miR159, miR167, miR398, miR408, and miR482 families (Zhai et al., 2011; Buxdorf et al., 2010;
Conclusions. Prediction lists comprising tens, hundreds, or more target sites for a single plant miRNA should be treated with extreme skepticism, as should meta-analyses of such lists (such as Gene Ontology enrichments). Such large numbers of targets are exceptional for a single plant miRNA, and there are strong theoretical and empirical arguments against the possibility of this occurring (Li et al., 2014b). If large numbers of targets for a single miRNA are found, there should be direct experimental evidence to support the assertions. Such exceptional cases do exist, such as the thousands of targets of the mir2118 family in rice that are experimentally supported by phasedRNA accumulation (Johnson et al., 2009). Those attempting to find plant miRNA targets should also be skeptical of inferring mechanistic attributes from the complementarity pattern. In particular, the psRNA-Target server (Dai and Zhao, 2011) marks miRNA-target alignments containing central mismatches as leading to translational repression. This contradicts a substantial amount of empirical data. Many plant miRNA targets that are perfectly paired in the central region are nonetheless also translationally repressed (Chen, 2004; Yang et al., 2012; Li et al., 2013). Conversely, not all targets that do have central mismatches are translationally repressed: Some are noncoding RNAs that are still sliced despite the central mismatches (Allen et al., 2009), some act as target mimics (Iwashita et al., 2011; Li et al., 2014), and others are simply nonfunctional, especially when the miRNA levels are not overwhelming (Li et al., 2014a).

Conservation, Evolution, and Annotations of Endogenous siRNAs

As listed above, the two prominent, endogenous siRNA populations in plants are phasedRNAs and hc-siRNAs. Both present unique challenges to annotation both because their sheer abundance can lead to appreciable false positives when annotating miRNAs and because of their differing patterns of conservation. PhasedRNAs are defined above and are mentioned here because they may be misidentified as miRNAs, due to their 21- or 22-nucleotide length, reproducibility across libraries and/or tissues, enrichment in specific tissues or cell types, and even their derivation from a processed inverted repeat. For example, rice miR5792 is derived from a locus that yields 24-nucleotide phasedRNAs only in meiotic stage anthers (Fei et al., 2016). Therefore, separation of miRNA candidates from phasedRNAs requires careful, often manual, assessment of each locus.

Endogenous hc-siRNAs comprise a major portion of the regulatory small RNA pool in key angiosperm model organisms including Arabidopsis (Lu et al., 2006), rice (Jeong et al., 2011), and maize (Zea mays; Nobuta et al., 2008). Most hc-siRNAs are 24 nucleotides in length and are thought to function in mRNA-directed DNA methylation, which couples 24-nucleotide siRNA production from heterochromatic regions to targeting of chromatin-associated nascent noncoding RNA transcripts. This targeting, which can be (and perhaps often is) non-cell autonomous (Melnyk et al., 2011) results in the de novo deposition of 5-methylcytosine at target loci (reviewed in Matzke et al., 2015). In Arabidopsis, complete removal of hc-siRNAs has only modest effects on genome-wide DNA methylation patterns (Stroud et al., 2013) and causes no obvious defects in organismal phenotype. In contrast, loss of hc-siRNAs in rice derepresses hundreds of protein-coding mRNAs, many of which are proximal to 24-nucleotide-generating miniature inverted repeat transposable elements (MITEs) (Wei et al., 2014). These features are relevant in the context of miRNA annotation because MITEs often contain miRNA-like inverted repeats, and heterochromatic regions may be low copy in the genome; hc-siRNAs from such regions may thus confound miRNA prediction algorithms and end up on output lists as candidate miRNAs. This is a particular challenge due to the somewhat circular argument that defining and annotating “heterochromatic” regions in a newly sequenced genome often depends on the presence of 24-nucleotide siRNAs.

sRNA-seq samples from several non-angiosperm plants lack an immediately obvious signature of abundant 24-nucleotide RNAs, including mosses (Axtell and Bartel,
COMMENTARY

2005; Arazi et al., 2005), lycophytes (Banks et al., 2008; Chávez Montes et al., 2014). These observations led to the suggestion that, unlike phasiRNAs and miRNAs, hc-siRNAs were not universal features found in all land plants (Dolgosheina et al., 2008). However, homologs of key genes known to be responsible for hc-siRNA biogenesis and function in angiosperms clearly exist in diverse plant lineages (Zong et al., 2009; Banks et al., 2011; Huang et al., 2015; Wang and Ma, 2015; You et al., 2017). Reverse genetic analyses of these homologs, coupled with sRNA-seq analyses of the mutants, has shown that hc-siRNAs exist in the moss P. patens (Cho et al., 2008; Coruh et al., 2015), which implies that the pathway was most likely present in the last common ancestor of all land plants. Whether hc-siRNAs have been specifically lost in the conifers and/or ferns remains an open question, but the presence of high levels of hc-siRNA, and phasiRNA annotations. Centralization of data, along with a single registry of locus names, will greatly facilitate future research, especially for those who are not small RNA specialists.

PROSPECTS

The two key developments in the study of plant regulatory RNA diversity and evolution are genome sequencing and assembly projects from diverse species, and high-throughput sRNA-seq data. We expect that the future will see continued acceleration in new plant genome assemblies and in accumulation of sRNA-seq data. As in many areas of biology, the challenge has shifted from acquisition of data to designing more robust and reproducible analyses of the data. Increasing the rigour of small RNA annotations, especially of hc-sRNAs and their targets, will be critical to prevent further degradation of central databases with overwhelming numbers of false positives. Increased use of transparent, easily reproducible sRNA-seq analytical methods should also act as a key community goal to ensure the deluge of small RNA data is put to good use.

ACKNOWLEDGMENTS

We thank Margaret Frank, Patricia Baldrich, and Matthew Endres for critical comments on this manuscript. We thank Seth Polydore for aligned sRNA-seq data. The U.S. National Science Foundation supports research in both authors’ labs (M.J.A.’s lab by award 1339207 and B.C.M.’s lab by awards 1257869 and 1339229).

AUTHOR CONTRIBUTIONS

M.J.A. and B.C.M. wrote the manuscript. M.J.A. analyzed data and prepared figures and tables.

REFERENCES


Chávez Montes, R.A., de Fátima Rosas-Cárdenas, F., De Paoli, E., Acrebi, M.,


The Plant Cell