Interactive Web-based Annotation of Plant MicroRNAs with iwa-miRNA

Ting Zhang[1,2], Jingjing Zhai[1,2], Xiaorong Zhang[1,2], Lei Ling[1,2], Menghan Li[3], Shang Xie[1,2], Minggui Song[4], Chuang Ma[1,2,*]

[1] *State Key Laboratory of Crop Stress Biology for Arid Areas, Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Shaanxi, Yangling 712100, China*

[2] *Key Laboratory of Biology and Genetics Improvement of Maize in Arid Area of Northwest Region, Ministry of Agriculture, Northwest A&F University, Shaanxi, Yangling 712100, China*

[3] *College of Plant Science, Tibet Agricultural and Animal Husbandry University, Tibet 860000, China*

[4] *College of Information Engineering, Northwest A&F University, Shaanxi, Yangling 712100, China*

* Corresponding author.

E-mail: chuangma2006@gmail.com; cma@nwafu.edu.cn (Ma C).

Running title: *Zhang T et al / iwa-miRNA: A Plant MicroRNA Annotation Tool*

Word counts: 5941

References: 54

Tables: 1

Figures: 5

Supplementary tables: 8

Counts of letters in the article title: 57

Keywords: 5

Words in Abstract: 201

Abstract

MicroRNAs (miRNAs) are important regulators of gene expression. The large-scale detection and profiling of miRNAs has accelerated with the development of high-throughput small RNA sequencing (sRNA-Seq) techniques and bioinformatics tools. However, generating high-quality comprehensive miRNA annotations remains challenging, due to the intrinsic complexity of sRNA-Seq data and inherent limitations of existing miRNA predictions. Here, we present iwa-miRNA, a Galaxy-based framework that can facilitate miRNA annotation in plant species by combining computational analysis and manual curation. iwa-miRNA is specifically designed to generate a comprehensive list of miRNA candidates, bridging the gap between already annotated miRNAs provided by public miRNA databases and new predictions from sRNA-Seq datasets. It can also assist users to select promising miRNA candidates in an interactive mode through the automated and manual steps, contributing to the accessibility and reproducibility of genome-wide miRNA annotation. iwa-miRNA is user-friendly and can be easily deployed as a web application for researchers without programming experience. With flexible, interactive, and easy-to-use features, iwa-miRNA is a valuable tool for annotation of miRNAs in plant species with reference genomes. We illustrated the application of iwa-miRNA for miRNA annotation of plant species with varying complexity. The sources codes and web server of iwa-miRNA is freely accessible at: http://iwa-miRNA.omicstudio.cloud/.

KEYWORDS: Galaxy; Interactive annotation; Manual inspection; MiRNA; Platform

Introduction

MicroRNAs (miRNAs) are a class of small non-coding RNAs that are widespread in eukaryotes and play roles in a variety of biological processes, including plant growth, development, and stress responses [1−3]. In plants, miRNA genes are transcribed into primary transcripts, which are processed by the DICER-LIKE 1 (DCL1), SERRATE (SE), and HYPONASTIC LEAVES 1 (HYL1) proteins to generate a stem-loop structured miRNA precursor, following by trimming into a mature miRNA/miRNA* duplex [4]. Recently, some

59 miRNAs have also been associated with agriculturally important traits, emerging as potential
60 targets for crop improvement and protection [5]. Due to their biological and agricultural
61 importance, miRNAs have become essential elements, annotated in genome sequences,
62 particularly for plant species.

63 Genome-wide miRNA identification is generally accomplished using bioinformatics
64 methods, such as homology search, machine learning-based prediction, and next-generation
65 sequencing (NGS) data mining [6−14]. Such computationally identified miRNAs from
66 multiple research groups have been deposited into public data repositories, such as miRBase
67 [15], PmiREN [16], sRNAanno [17], and Plant small RNA genes (PsRNA) [18], providing
68 valuable resources for life scientists interested in miRNA research, —from single-gene to
69 genome-wide scale, basic molecular biology to population genetics, and bioinformatics to
70 experimental biology. Despite these advances, present-day annotations remain riddled with
71 false positives, and have a limited degree of comprehensiveness (the fraction of all *bona fide*
72 miRNA genes that are included), exhaustiveness (the fraction of all mature miRNAs from
73 each miRNA gene), and completeness (the fraction of pri/pre-miRNA sequences that cover
74 the entire length) [19,20].

75 There are multiple factors that make the computational identification of miRNAs
76 challenging. First, the tissue-/cell-type/developmental-stage specific expression and/or low
77 expression properties of some miRNAs mean that they are often poorly identified from
78 traditional low-throughput experimental studies and NGS experiments with limited samples
79 and/or low sequencing depth. Second, the imperfect criteria defined for identification of
80 miRNAs from NGS data. Although high-throughput criteria were established years ago [21],
81 and have been continuously updated in response to studies of mechanisms [12,19], they
82 cannot fully capture the species to species variation of miRNA characteristics. Third, the
83 unsatisfactory level of accuracy of automatic miRNA annotation methods. Homology-based
84 strategies fail to identify species-specific miRNAs, while machine learning-based tools have
85 been designed for genome-wide miRNA prediction [7,8,22−24]; however, most were trained
86 with limited experimentally validated miRNA data, and have markedly lower prediction
87 accuracy in cross-species prediction [25]. Since the introduction of small RNA sequencing
88 (sRNA-Seq), many sequencing-based tools have been developed that vary in their

characterization of miRNAs [10−14]; however, only few tools have kept pace with updated miRNA identification criteria and they continue to suffer from trade-offs between quality and quantity [26]. Given these differences in the use of sRNA-Seq data, automatic annotation approaches, and miRNA identification criteria, inconsistency often arises in existing plant miRNA annotations. For example, at the beginning of this project in December 2019, we observed that, in *Arabidopsis thaliana*, there were 326, 221, 163, and 142 miRNA precursors annotated in miRBase (v22.1), PmiREN (v1.0), sRNAanno (v1.0), and PsRNA (v1.0) databases, respectively, with an overlap of only 120 miRNA precursors among these four databases. These inconsistencies indicate a proportion of false positives and false negatives within existing plant miRNA annotations, using which may result in unappreciated, but often serious, consequences for downstream studies, such as expression quantification, differential expression analysis, targetome analysis, and functional screening.

A straightforward way to improve the quality of miRNA annotations is to develop bioinformatics methods with sophisticated design of miRNA identification algorithms and criteria. In addition, a combination of automatic annotation and manual annotation would also be effective. The power of manual annotation has been demonstrated in human protein-coding gene annotation, where a team of human annotators inspect evidence supporting automatically annotated transcripts, and create relatively confident annotation databases, including GENCODE [27] and Ref-Seq [28]. These manual annotation databases are often free from many of the artefacts inherent in automated approaches, and have been adopted by most large-scale genomics projects, including the Encyclopedia of DNA Elements (ENCODE) [29] and the Genotype-Tissue Expression (GTEx) project [30]. In recent years, manual inspection has also been advocated, and performed to compile high-quality miRNA dataset from the genomes of human [31] and *Citrus sinensis* [26]. These pioneer investigations will provoke a wider interest among scientists in the research field of manual inspection of genome annotation, accompanied by an increased demand for effective interactive annotation tools to manage and analyze genome-wide miRNA annotations.

Here, we present iwa-miRNA, a web-based framework for interactive annotation of miRNAs from plant species with reference genomes. iwa-miRNA not only provides functions for automatically incorporating miRNA annotations from four representative databases (i.e.,

miRBase, PmiREN, sRNAanno, and PsRNA), but also builds a bioinformatics pipeline designed specifically to handle large-scale sRNA-Seq data for candidate miRNA prediction. Both annotated and predicted miRNAs are aggregated into a comprehensive list of miRNA candidates. Two miRNA selection approaches, high-throughput criteria and machine learning-based, are provided to assist the selection of promising miRNA candidates, based on the sequence-, structure-, and expression-based features. To enhance the accessibility of miRNA annotation, iwa-miRNA generates a report page with detailed information customized by feature types for each selected miRNA, facilitating convenient miRNA refinement during manual curation. The source codes of iwa-miRNA have been combined into a Galaxy platform, which are organized with a user-friendly web interface, and supported with extensive user documents. With these flexible, interactive, easy-to-use features, iwa-miRNA can generate a comprehensive collection of miRNA candidates and allows users to interrogate miRNA annotation in a straightforward way, without the need for computational skills. We provide examples of the application of iwa-miRNA for miRNA annotation of *Arabidopsis thaliana*, maize (*Zea mays* L.), and hexaploid bread wheat (*Triticum aestivum* L.).

Method

The iwa-miRNA comprises three modules: ***MiRNA Compilation, MiRNA Selection,*** and ***Manual Curation*** (**Figure 1; Table 1**). The source codes of the modules and their dependencies are fully organized within the Galaxy framework. iwa-miRNA can be expertly implemented through a user-friendly web interface, and summarizes the results into HTML pages, using Rmarkdown for easy visualization, interpretation, and sharing.

MiRNA Compilation (Module I): This module generates a comprehensive collection of miRNA candidates by aggregating already annotated miRNAs from four plant miRNA databases (i.e., miRBase, PmiREN, sRNAanno, and PsRNA) and predicted miRNAs from user-submitted sRNA-Seq data (Figure 1A). For a plant species of interest, all miRNA annotations (e.g. name, sequence, genomic coordinates, and so on) provided by the four miRNA databases are automatically retrieved using the “miRNARetrival” function. miRNA

prediction is accomplished using the "miRNAPredict" function, which is specifically designed for parallel analysis of large-volume sRNA-Seq data. This function is based on a series of bioinformatics tools and custom scripts required for read cleaning (FASTX-Toolkit v0.0.14; http://hannonlab.cshl.edu/fastx_toolkit), genome mapping of reads (Bowtie v1.2.3 [32]), and miRNA prediction (miRDeep-P2 [12] and miRCat2 [33]). iwa-miRNA accepts the inputs of target genome sequences in FASTA format and corresponding annotations in GFF3/GTF format, which can be directly submitted by users or automatically fetched from the Ensembl Plants (https://plants.ensembl.org) database using the 'genomePrepare' tool. For miRNA annotations from different versions of the genome, the 'miRNATranslate' function can be used to translate annotated miRNAs into the genomic coordinate system of the target genome by performing miRNA precursor-to-genomic alignment using GMAP (v2019.09.12) [34]. All miRNA candidates are finally organized using a uniform naming scheme and genomic coordinates.

MiRNA Selection (Module II): This module selects a subset of miRNA candidates that are regarded as promising miRNAs, according to the high-throughput criteria and/or using a machine learning-based approach (Figure 1B). For the latter miRNA selection approach, iwa-miRNA builds a one-class support vector machine (SVM) classifier [35] to predict if tested miRNA candidates are potentially real miRNAs or not (see File S1 and Table S1 for details). iwa-miRNA is user friendly, in that users can tune corresponding parameters according to the sRNA-Seq data at hand. A set of default parameters, derived from our own analysis experience, are also provided to assist non-expert users within their analyses.

Manual Curation (Module III): This module provides the information for all miRNA candidates generated during the compilation and selection processes, and creates a summary page for rapid curation of the quality of selected miRNAs (Figure 1C). For miRNAs of interest, users can further inspect them by entering into the corresponding report pages, which provides more detailed information customized by feature types and visualized using various plot styles. A secondary structure plot is generated to display the location of a mature miRNA within the precursor sequence and quality-profiling results. Read stacks are plotted to show the read support of identified miRNAs. A boxplot is used to visualize miRNA expression patterns and arm selection events across different samples. A bipartite network is constructed

178 to depict miRNA-target interactions predicted by psRNAtarget [36]. Users can update the list
179 of selected miRNAs in a dynamic manner through adjusting criteria thresholds, or by direct
180 deletion from the summary page. Selected miRNAs are finally exported into table,
181 GFF3/GTF, and FASTA format files for downstream exploratory analysis (Figure 1D).
182

183 ## Results

184 We illustrate the efficiency of iwa-miRNA for miRNA annotation of three plant genomes of
185 different complexity. Among these applications, four databases (miRBase, PmiREN,
186 sRNAanno, and PsRNA) and a set of publicly available sRNA-Seq datasets were used to
187 generate candidate miRNAs. Both high-throughput criteria and one-class SVM with default
188 parameters (Figure 1B) were used to identify promising miRNA candidates.
189

190 **Cases 1: Application of iwa-miRNA for miRNA annotation in *Arabidopsis***

191 As an initial demonstration of our framework, we looked at the long-studied and extensively
192 annotated miRNAs of the model plant species, *Arabidopsis*, which has a relatively small
193 genome of ~135 Mb. In *Arabidopsis*, miRNAs have been studied for over 18 years [37−39],
194 and explored using more than 2000 sRNA-Seq datasets [40]. Using iwa-miRNA, we obtained
195 a total of 365 miRNA precursors corresponding to 625 mature forms, from the four databases
196 (miRBase, PmiREN, sRNAanno, and PsRNA; **Figure 2**A; Table S2). Using 1063 sRNA-Seq
197 datasets (see details in File S1) for Columbia ecotype of *Arabidopsis* as inputs, iwa-miRNA
198 predicted 435 miRNA precursors, 302 of which were not previously annotated in any of the
199 four plant miRNA databases. This resulted in generation of 667 miRNA precursor candidates,
200 corresponding to 1190 mature miRNA candidates for *Arabidopsis* (Table S2).

201 Newly predicted miRNA precursors were expressed at remarkably lower expression
202 levels and with less breath than those that were already annotated (Figure 2B−2C), indicating
203 the potential importance of iwa-miRNA in identifying miRNA precursors with
204 tissue-/cell-type/developmental-stage specific expression and/or low expression patterns.
205 There were 330 miRNA precursor candidates that passed the high-throughput criteria
206 (denoted as Ara-Set1), 203 of which were annotated in at least one of the four databases.

Using iwa-miRNA, we were able to characterize these 667 candidate miRNA precursors using 219 sequence features, 382 structural features, and 1,063 expression features (i.e., transcripts per million [TPM] values across all samples), providing an opportunity to predict miRNAs using machine learning approaches (see File S1 for details). Using 365 already annotated miRNA precursors as positive samples, iwa-miRNA built a one-class SVM classifier to predict 308 miRNA precursor candidates (Ara-Set2) as true positives. There were 208 candidate miRNA precursors in common between Ara-Set1 and Ara-Set2 (Figure 2D). For newly predicted miRNA precursors, 15 candidates (five from each region of the Venn diagram at the bottom of Figure 2D) were randomly selected for validation using quantitative real-time polymerase chain reaction (qRT-PCR) experiments, in which mature sequences were amplified with specifically designed primers (Table S3). Three miRNA precursor candidates were excluded because their mature sequences were not unique in the *Arabidopsis* reference genome sequence. These wet laboratory experiments validated 11 candidates as expressed in a mixed sample of *Arabidopsis* roots, shoots, leaves, and flowers (Figure 2E; File S1). These results provide evidence to confidently annotate *Arabidopsis* miRNAs using iwa-miRNA, although further validation experiments at large scale are desirable.

For each miRNA precursor candidate, iwa-miRNA assigns an identifier *via* a uniform naming scheme and the corresponding uniform resource identifier (URI) is hyperlinked to an HTML web page reporting a detailed description of feature information for manual inspection (**Figure 3**A). Figure 3B shows the report page of a representative example "ath-MIR156b", which regulates vegetative phase change and recurring environmental stress by repressing squamosa promoter binding protein-like (SPL) transcription factors [41,42]. The precursor of miR156b produces two mature miRNAs of different lengths: a 20-nt miRNA from the 5' arm (miR156b-5p) and a 23-nt miRNA from the 3' arm (miR156b-3p). The former is high levels in root, leaf, and seed tissues, while the latter is preferentially expressed in root.

Cases 2: Application of iwa-miRNA for miRNA annotation in maize

The successful application of iwa-miRNA to miRNA annotation in *Arabidopsis* prompted us to evaluate its value for analysis of plants with larger, more complex genomes. Here we

237 focused on the model crop, maize, specifically the B73 inbred line, which has a reference
238 genome of 2.3 Gb, more than 80% of which comprises transposable elements and other
239 repeat sequences [43]. iwa-miRNA obtained a total of 619 miRNA precursors, which
240 correspond to 893 mature forms, from the four databases (**Figure 4**A). For each database, the
241 proportion of uniquely annotated maize miRNA precursors was markedly different from that
242 observed in *Arabidopsis* (Table S4). This difference underscores the importance of
243 performing an aggregation analysis and manual inspection of miRNA annotations from
244 different sources. Further, on integrative analysis of 195 sRNA-seq datasets obtained from
245 previously reported work [44], a total of 1241 miRNA precursor candidates were identified,
246 622 of which were previously un-annotated in any of the four plant miRNA databases (Table
247 S5).

248 Using iwa-miRNA, 886 miRNA precursors were selected for downstream analysis: 588
249 passed with the high-throughput criteria (maize-Set1) and 704 predicted by the one-class
250 SVM classifier (maize-Set2) (Figure 4B). One of the SVM-predicted miRNAs,
251 zma-miR_N85a-5p, had already been validated by qRT-PCR in our recently published paper
252 [44], and exhibited a specifical expression pattern in seed tissues. Of these 886 miRNA
253 precursors, 381 exhibited broad expression patterns, having $\geqslant$ 1 TPM in more than 50% of
254 195 samples (Figure 4C).

255 Preliminary statistical analysis showed that most novel miRNAs were 21-nt and 22-nt in
256 length (Table S5). Novel miRNAs were predominantly from intergenic regions and
257 transposable elements (Figure 4D). Some newly predicted miRNAs are statistically
258 associated with maize traits. As shown in Figure 4E, in the tail region of zma-miR_N221a-5p,
259 there is a single nucleotide polymorphism (SNP; rs#chr6: 22134053) that has been reported
260 to be significantly associated with the metabolic trait
261 "Np.Npp_Feruloyl.caffeoyl_spermidine_derivative_E1" (*P*-value = 4.7E-6) [45]. This
262 genetic variant (AA/GG) may influence target gene selection (56 *vs*. 51 genes; 30
263 overlapped). These preliminary results indicate the efficiency and power of interactive
264 annotation of miRNAs in crop plants with complex genomes.

265

266 **Cases 3: Application of iwa-miRNA for miRNA annotation in wheat**

Finally, we showcased the application of iwa-miRNA to miRNA annotation in a more complex crop plant, hexaploid (AABBDD) bread wheat (*Triticum aestivum*), which has an even larger genome (~17 Gb), with a more complex repeat landscape, than that of maize [46]. In addition to this, the wheat genome also presents other specific challenges, such as the composition of three closely-related and independently maintained subgenomes. Given that miRNAs were previously annotated based on different versions of the wheat reference genome from the four plant miRNA databases, iwa-miRNA first unified the miRNA annotation by mapping miRNA precursor sequences to the latest wheat reference genome (IWGSC RefSeq v1.0; https://plants.ensembl.org/Triticum_aestivum) using GMAP (v2019.09.12). Then, it was applied to predict miRNA precursors from 95 sRNA-Seq datasets (see details in File S1; Table S6), resulting in a list of 2857 miRNA precursor candidates in wheat (**Figure 5**A; Table S7). Subsequently, 2030 miRNA precursor candidates were selected, based on high-throughput criteria (wheat-Set1; 1617 miRNA precursor candidates) and one-class SVM prediction (wheat-Set2; 1410 miRNA precursor candidates) (Figure 5B). Finally, these 2030 selected miRNA precursors, corresponding to 2163 mature miRNAs, were organized into a summary page for future experimental validation and functional exploration.

Of 2030 miRNA precursors, 1926 were clearly located on the three subgenomes, among which the D subgenome contained significantly fewer miRNA precursors than the A and B subgenomes (571 *vs*. 611 and 744, respectively; χ^2 test, $P < 0.001$). This result suggests that there may be subgenome bias within these annotated miRNAs. Of 1926 miRNA precursors (A: 611; B: 744; D: 571), 1919 formed 1565 homologous groups of eight A:B:D configurations: 1:1:1 (118), 1:1:0 (23), 1:0:1 (33), 0:1:1 (31), 1:0:0 (417), 0:1:0 (561), 0:0:1 (370), and others (12) (Figure 5C; Table S8). Further, 7.54% (118/1565) of groups of homologous miRNA genes contained triads, with a single gene copy per subgenome. Among these 118 triads, 17 groups exhibited differences in mature sequences and expression levels (Figure 5D). A representative example is Tae-MIR408a/c/f, which has been linked to salt stress response and leaf polarity in wheat [5,47,48]. Among the A, B, and D subgenomes, there is a single nucleotide difference exists in the Tae-MIR408a/c/f 5' mature sequence (A/B *vs*. D) and the 3' mature sequence (A/D *vs*. B), resulting in identification of four mature

297 sequences: Tae-miR408a-5p, Tae-miR408c/f-5p, Tae-miR408a/c-3p, and Tae-miR408f-3p
298 (Figure 5E). These four sequences exhibit different expression patterns in leaf and grain
299 tissues (Figure 5F). Tae-miR408a-5p and Tae-miR408c/f-5p were highly expressed in grain
300 and leaf, respectively. Further, Tae-miR408a/c-3p was mainly expressed in leaf, while
301 Tae-miR408f-3p had no expression in any tissues tested. A nucleotide difference in mature
302 sequences also results in diversity in target genes (Figure 5G). These results indicate that the
303 comprehensive annotation of miRNAs provides opportunities to explore evolution and
304 functional diversification of miRNAs in polyploid plants, including hexaploid bread wheat.
305

306 ## Discussion

307 MiRNAs have been the subject of extensive research over the past 20 years [39,49]; however,
308 annotating miRNAs at the genome scale is not straightforward, not only because of the
309 experimental difficulty in capturing some miRNAs with low- or context-specific expression
310 patterns, but also owning to the computational difficulties in accurately distinguishing signals
311 from noises within genomic sequences and sRNA-Seq data. To facilitate miRNA annotation,
312 we have developed an interactive annotation framework, iwa-miRNA, with a user-friendly
313 interface, to compile, select, and manually curate miRNAs from plant species with reference
314 genomes.

315 Compared with existing miRNA-related bioinformatics annotation databases and tools,
316 iwa-miRNA has several distinguishing features.

317 First, iwa-miRNA bridges the gap between existing annotations provided by public
318 miRNA databases and new predictions from sRNA-Seq datasets. Most miRNA databases are
319 periodically updated; however, they cannot integrate new miRNAs predicted from the rapidly
320 accumulating sRNA-Seq data in a timely manner. In contrast, many bioinformatics tools have
321 been designed with the sole purpose of predicting miRNAs from sRNA-Seq data, with less
322 consideration of miRNA annotations provided by different databases. This issue was recently
323 addressed by miRCarta [50], which collects novel human miRNA candidates and augments
324 the annotation information provided by miRBase. Unlike miRCarta, iwa-miRNA tackles this
325 deficiency with an emphasis on plant miRNAs. We suggest that, in the future, more attention
326 should be paid to bridging the gap between miRNA annotations and predictions in plants,

human, and other species.

Second, iwa-miRNA provides a new way to interactively select promising miRNAs. The aggregation of already annotated and newly predicted miRNAs generates a comprehensive collection of miRNA candidates, which certainly contain false positive hits, as well as interesting candidates (especially those expressed in tissue-/cell-type/developmental-stage specific patterns) for further validation. In previous studies, miRNA candidates were selected according to pre-defined rules, with 'black box' type functionality. The reliability of computationally annotated miRNAs is difficult to judge, particularly for scientists who are not involved in the process of miRNA prediction and annotation. iwa-miRNA provides a visualization of the sequence-, structure-, and expression-based features used in miRNA selection. In this way, with accessible miRNA annotation, researchers (both annotators and bench scientists) can manage miRNA selection in a dynamic manner, thus conveniently selecting promising candidates for further exploratory analysis and experimental validation.

Third, iwa-miRNA is user-friendly and can be deployed as a web application for easy accessibility and public/private data analysis. To facilitate the application of iwa-miRNA, all source codes and dependencies have been integrated into the Galaxy system, followed by packaging an independent runtime environment into a Docker image. This enables compatibility and portability: users can launch iwa-miRNA using a simple command, regardless of which operation system (Windows, Linux, or Macintosh), is used. Through a simple graphical interface, users can use this 'one-stop' systematic platform to mine massively accessible sRNA-Seq datasets and miRNA annotation. iwa-miRNA is also integrated with an Rmarkdown-based HTML report, to return interactive tables, publication-grade plots, and reproducible operations.

iwa-miRNA also suffers from some limitations. It cannot handle sRNA-Seq data from species without genome sequences. Recent RNA-Seq data analysis revealed that unmapped reads can serve as a valuable resource for new gene discovery [51]. In the current version, iwa-miRNA ignores sequencing reads that do not map to the reference genome. Since the main purpose of this study was to develop a platform that facilitates integrative annotation of miRNAs, iwa-miRNA has limited ability to perform downstream analysis of selected

357 miRNAs, such as enrichment analysis (e.g. microRNA gene ontology annotations and
358 miRNA set enrichment analysis) and comparative analysis between different species [52−54].
359 It should be also noted that there may be some false positives contained in the selected list of
360 miRNA candidates. Further "wet" experiments should be performed to validate miRNA
361 candidates of particular interest before any functional investigation.
362

363 ## Future work

364 The iwa-miRNA project continues to be developed and improved. In future versions of
365 iwa-miRNA, we plan to develop new functional modules to analyze sRNA-Seq data for
366 species without a reference genome, identify candidate miRNAs from unmapped reads, and
367 provide more downstream exploratory analysis. We invite researchers to use the iwa-miRNA
368 platform to carry out large-scale sRNA-Seq data analysis on their local computers. Research
369 collaborations are welcome, in particular for researchers without high-throughput
370 computational resources.
371

372 ## Data availability

373 iwa-miRNA Docker image is available at https://hub.docker.com/r/malab/iwa-mirna. Source
374 codes and user manual are available at https://github.com/cma2015/iwa-miRNA. The
375 prototype web server of iwa-miRNA is accessible at http://iwa-miRNA.omicstudio.cloud/.
376

377 ## CRediT authorship contribution statement

378 **Ting Zhang**: Methodology, Software, Formal analysis, Data curation, Visualization, Writing
379 - original draft, Writing - review & editing. **Jingjing Zhai**: Methodology, Software, Formal
380 analysis, Writing - review & editing. **Xiaorong Zhang**: Data curation, Software. **Lei Ling**:
381 Software, Formal analysis. **Menghan Li**: Data curation. **Shang Xie**: Software. **Minggui**
382 **Song**: Software, Visualization. **Chuang Ma**: Conceptualization, Methodology, Supervision,
383 Funding acquisition, Project administration, Writing - original draft, Writing - review &
384 editing. All authors read and approved the final manuscript.
385

Competing interests

The authors have declared no competing interests.

Acknowledgements

We thank High-Performance Computing (HPC) of Northwest A&F University for providing computing resources. This work has been supported by the National Natural Science Foundation of China (31570371), the Youth 1000-Talent Program of China, the Hundred Talents Program of Shaanxi Province of China, and the Fundamental Research Funds for the Central Universities (2452020041).

Authors' ORCID IDs

0000-0002-1360-7683 (Ting Zhang)

0000-0002-1535-3103 (Jingjing Zhai)

0000-0001-6916-4407 (Xiaorong Zhang)

0000-0001-6426-533X (Lei Ling)

0000-0001-8038-8559 (Menghan Li)

0000-0002-9751-1418 (Shang Xie)

0000-0002-4406-3877 (Minggui Song)

0000-0001-9612-7898 (Chuang Ma)

406 # References

407 [1] Yu Y, Jia T, Chen X. The 'how' and 'where' of plant microRNAs. New Phytol
408 2017;216:1002−17.
409 [2] D'Ario M, Griffiths-Jones S, Kim M. Small RNAs: big impact on plant development.
410 Trends Plant Sci 2017;22:1056−68.
411 [3] Moran Y, Agron M, Praher D, Technau U. The evolutionary origin of plant and animal
412 microRNAs. Nat Ecol Evol 2017;1:27.
413 [4] Voinnet O. Origin, biogenesis, and activity of plant microRNAs. Cell 2009;136:669−87.
414 [5] Tang J, Chu C. MicroRNAs in crop improvement: fine-tuners for complex traits. Nat
415 Plants 2017;3:17077.
416 [6] Artzi S, Kiezun A, Shomron N. miRNAminer: a tool for homologous microRNA gene
417 search. BMC Bioinformatics 2008;9:39.
418 [7] Meng J, Liu D, Sun C, Luan Y. Prediction of plant pre-microRNAs and their microRNAs
419 in genome-scale sequences using structure-sequence features and support vector machine.
420 BMC Bioinformatics 2014;15:423.
421 [8] Cui H, Zhai J, Ma C. miRLocator: machine learning-based prediction of mature
422 microRNAs within plant pre-miRNA sequences. PLoS One 2015;10:e0142753.
423 [9] Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes.
424 RNA 2013;19:740−51.
425 [10] Lei J, Sun Y. miR-PREFeR: an accurate, fast, and easy-to-use plant miRNA prediction
426 tool using small RNA-Seq data. Bioinformatics 2014;30:2837−9.
427 [11] Evers M, Huttner M, Dueck A, Meister G, Engelmann JC. miRA: adaptable novel
428 miRNA identification in plants using small RNA sequencing data. BMC Bioinformatics
429 2015;16:370.
430 [12] Kuang Z, Wang Y, Li L, Yang X. miRDeep-P2: accurate and fast analysis of the
431 microRNA transcriptome in plants. Bioinformatics 2019;35:2521−2.
432 [13] Aparicio-Puerta E, Lebron R, Rueda A, Gomez-Martin C, Giannoukakos S, Jaspez D, et
433 al. sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential
434 expression. Nucleic Acids Res 2019;47:W530−5.

435 [14] Liu Q, Ding C, Lang X, Guo G, Chen J, Su X. Small noncoding RNA discovery and
436 profiling with sRNAtools based on high-throughput sequencing. Brief Bioinform 2019;
437 bbz151.
438 [15] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to
439 function. Nucleic Acids Res 2019;47:D155−62.
440 [16] Guo Z, Kuang Z, Wang Y, Zhao Y, Tao Y, Cheng C, et al. PmiREN: a comprehensive
441 encyclopedia of plant miRNAs. Nucleic Acids Res 2020;48:D1114−21.
442 [17] Chen C, Feng J, Liu B, Li J, Feng L, Yu X, et al. sRNAanno—a database repository of
443 uniformly-annotated small RNAs in plants. bioRxiv 2019:771121.
444 [18] Lunardon A, Johnson NR, Hagerott E, Phifer T, Polydore S, Coruh C, et al. Integrated
445 annotations and analyses of small RNA-producing loci from 47 diverse plants. Genome Res
446 2020;30:497−513.
447 [19] Axtell MJ, Meyers BC. Revisiting criteria for plant microRNA annotation in the era of
448 big data. Plant Cell 2018;30:272−84.
449 [20] Alles J, Fehlmann T, Fischer U, Backes C, Galata V, Minet M, et al. An estimate of the
450 total number of true human miRNAs. Nucleic Acids Res 2019;47:3353−64.
451 [21] Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, et al. Criteria
452 for annotation of plant MicroRNAs. Plant Cell 2008;20:3186−90.
453 [22] Morgado L, Johannes F. Computational tools for plant small RNA detection and
454 categorization. Brief Bioinform 2019;20:1181−92.
455 [23] Stegmayer G, Di Persia LE, Rubiolo M, Gerard M, Pividori M, Yones C, et al.
456 Predicting novel microRNA: a comprehensive comparison of machine learning approaches.
457 Brief Bioinform 2019;20:1607−20.
458 [24] Chen L, Heikkinen L, Wang C, Yang Y, Sun H, Wong G. Trends in the development of
459 miRNA bioinformatics tools. Brief Bioinform 2019;20:1836−52.
460 [25] Leclercq M, Diallo AB, Blanchette M. Computational prediction of the localization of
461 microRNAs within their pre-miRNA. Nucleic Acids Res 2013;41:7200−11.
462 [26] Taylor RS, Tarver JE, Foroozani A, Donoghue PC. MicroRNA annotation of plant
463 genomes − Do it right or not at all. Bioessays 2017;39.

464 [27] Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, et al. GENCODE:
465 producing a reference annotation for ENCODE. Genome Biol 2006;7 Suppl 1:S4.1−9.
466 [28] Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq:
467 an update on prokaryotic genome annotation and curation. Nucleic Acids Res
468 2018;46:D851−60.
469 [29] Consortium EP. An integrated encyclopedia of DNA elements in the human genome.
470 Nature 2012;489:57−74.
471 [30] Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet
472 2013;45:580−5.
473 [31] Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, et al. A uniform system
474 for the annotation of vertebrate microRNA genes and the evolution of the human
475 microRNAome. Annu Rev Genet 2015;49:213−42.
476 [32] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient
477 alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25.
478 [33] Paicu C, Mohorianu I, Stocks M, Xu P, Coince A, Billmeier M, et al. miRCat2: accurate
479 prediction of plant and animal microRNAs from next-generation sequencing datasets.
480 Bioinformatics 2017;33:2446−54.
481 [34] Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA
482 and EST sequences. Bioinformatics 2005;21:1859−75.
483 [35] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM transactions
484 on intelligent systems and technology (TIST) 2011;2:1−27.
485 [36] Dai X, Zhuang Z, Zhao PX. psRNATarget: a plant small RNA target analysis server
486 (2017 release). Nucleic Acids Res 2018;46:W49−54.
487 [37] Llave C, Kasschau KD, Rector MA, Carrington JC. Endogenous and
488 silencing-associated small RNAs in plants. Plant Cell 2002;14:1605−19.
489 [38] Park W, Li J, Song R, Messing J, Chen X. CARPEL FACTORY, a Dicer homolog, and
490 HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. Curr Biol
491 2002;12:1484−95.
492 [39] Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP. MicroRNAs in plants.
493 Genes Dev 2002;16:1616−26.

494 [40] Feng L, Zhang F, Zhang H, Zhao Y, Meyers BC, Zhai J. An online database for
495 exploring over 2,000 *Arabidopsis* small RNA libraries. Plant Physiol 2020;182:685−91.
496 [41] Wu G, Park MY, Conway SR, Wang JW, Weigel D, Poethig RS. The sequential action
497 of miR156 and miR172 regulates developmental timing in *Arabidopsis*. Cell
498 2009;138:750−9.
499 [42] Stief A, Altmann S, Hoffmann K, Pant BD, Scheible WR, Baurle I. *Arabidopsis* miR156
500 regulates tolerance to recurring environmental stress through SPL Transcription Factors.
501 Plant Cell 2014;26:1792−807.
502 [43] Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference
503 genome with single-molecule technologies. Nature 2017;546:524−7.
504 [44] Xu Y, Zhang T, Li Y, Miao Z. Integrated analysis of large-scale omics data revealed
505 relationship between tissue specificity and evolutionary dynamics of small RNAs in maize
506 (*Zea mays*). Front Genet 2020;11:51.
507 [45] Gui S, Yang L, Li J, Luo J, Xu X, Yuan J, et al. ZEAMAP, a comprehensive database
508 adapted to the maize multi-omics era. iScience 2020;23:101241.
509 [46] International Wheat Genome Sequencing C, investigators IRp, Appels R, Eversole K,
510 Feuillet C, Keller B, et al. Shifting the limits in wheat research and breeding using a fully
511 annotated reference genome. Science 2018;361:eaar7191.
512 [47] Feng H, Zhang Q, Wang Q, Wang X, Liu J, Li M, et al. Target of tae-miR408, a
513 chemocyanin-like protein gene (*TaCLP1*), plays positive roles in wheat response to
514 high-salinity, heavy cupric stress, and stripe rust. Plant Mol Biol 2013;83:433−43.
515 [48] Zhao XY, Hong P, Wu JY, Chen XB, Ye XG, Pan YY, et al. The tae-miR408-mediated
516 control of *TaTOC1* genes transcription is required for the regulation of heading time in wheat.
517 Plant Physiol 2016;170:1578−94.
518 [49] Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes
519 small RNAs with antisense complementarity to *lin-14*. Cell 1993;75:843−54.
520 [50] Backes C, Fehlmann T, Kern F, Kehl T, Lenhof HP, Meese E, et al. miRCarta: a central
521 repository for collecting miRNA candidates. Nucleic Acids Res 2018;46:D160−7.
522 [51] Chen S, Ren C, Zhai J, Yu J, Zhao X, Li Z, et al. CAFU: a galaxy framework for
523 exploring unmapped RNA-Seq data. Brief Bioinform 2020;21:676−86.

524 [52] Huntley RP, Sitnikov D, Orlic-Milacic M, Balakrishnan R, D'Eustachio P, Gillespie ME,
525 et al. Guidelines for the functional annotation of microRNAs using the gene ontology. RNA
526 2016;22:667−76.
527 [53] Kern F, Fehlmann T, Solomon J, Schwed L, Grammes N, Backes C, et al. miEAA 2.0:
528 integrating multi-species microRNA enrichment analysis and workflow management systems.
529 Nucleic Acids Res 2020;48:W521-8.
530 [54] Gramzow L, Theissen G. Plant miRNA conservation and evolution. Methods Mol Biol
531 2019;1932:41−50.
532

533 ## Figure legends

534 **Figure 1 Graphical summary of iwa-miRNA**

535 **A.** MiRNA candidates are generated by aggregating annotated and predicted miRNAs. B.
536 Promising miRNAs are selected using high-throughput criteria and machine learning
537 approaches. **C.** Manual curation of selected miRNAs based on annotation information from
538 summary and report pages. **D.** Exploratory analysis of selected miRNAs. PCGs, protein
539 coding genes; TEs, transposable elements; TIR, terminal inverted repeat; LINE, long
540 interspersed element; LTR, long terminal repeat.
541

542 **Figure 2 Application of miRNA annotation in *Arabidopsis***

543 **A.** Venn diagram comparing miRNA precursors provided by four miRNA databases. **B.**
544 Cumulative frequency of log2 expression levels of already annotated and newly predicted
545 miRNA precursor candidates. **C.** Number of expressed miRNA precursor candidates (already
546 annotated and newly predicted) in different samples. **D.** Venn diagram comparing miRNA
547 precursor candidates between the two miRNA selection approaches. **E.** Quantitative RT-PCR
548 (qRT-PCR) results for 11 candidates. CT values were normalized to U6 small RNA. All
549 values represent means ± SE (n = 3). TPM, transcripts per million; SVM, support vector
550 machine; CT, cycle threshold.
551

552 **Figure 3 Summary and report pages generated by iwa-miRNA**

A. Screenshot of summary page reporting the information of some features for *Arabidopsis* miRNAs. **B.** The report page for ath-MIR156b. HT, high-throughput; MFE, minimal free energy; AMFE, adjusted minimal free energy.

Figure 4 Application of miRNA annotation in maize

A. Venn diagram comparing miRNA precursors provided by four miRNA databases. **B.** Venn diagram comparing miRNA precursor candidates generated by two miRNA selection approaches. **C.** Expression levels of 886 miRNA precursors in 195 samples. Adjacent bar chart indicates three categorical results. The right panel shows the percentage of samples with expression greater than different thresholds (e.g. 1, 5, 10, etc.) **D.** Distribution of different miRNA lengths among different genomic features. **E.** The effect of SNP rs#chr6:22134053 on target genes of zma-MIR_N221a.

Figure 5 Application of miRNA annotation in wheat

A. Venn diagram comparing miRNA precursors provided by four miRNA databases. **B.** Venn diagram comparing miRNA precursor candidates generated by two miRNA selection approaches. **C.** Number of homologous groups and miRNAs in different A:B:D configurations. **D.** Ternary plot of miRNA expression levels from A, B, and D in triads. Tae-MIR408a/c/f is indicated by a red arrow. **E.** Schematic diagram of Tae-MIR408a/c/f in the ABD subgenomes. **F.** Expression levels of four mature sequences of Tae-MIR408a/c/f in leaf and grain tissues. **G.** The nucleotide difference and target alteration between mature sequences of Tae-MIR408a/c/f.

Tables

Table 1 Overview of functional modules in iwa-miRNA

Supplementary material

File S1 Detailed information regarding sRNA-Seq data collection and preprocess, SVM modeling, qRT-PCR experiment, and syntenic analysis of the wheat genome

582 **Table S1 Sequence and structural features used in iwa-miRNA**

583 **Table S2 List of candidate miRNAs and their precursors in *Arabidopsis***

584 **Table S3 Primers for miRNA qRT-PCR assay in *Arabidopsis***

585 **Table S4 Proportion of miRNA precursors uniquely annotated in each of four miRNA**

586 **databases**

587 **Table S5 List of candidate miRNAs and their precursors in maize**

588 **Table S6 List of 95 wheat sRNA-Seq libraries**

589 **Table S7 List of candidate miRNAs and their precursors in wheat**

590 **Table S8 Homologous groups of A:B:D configurations and detailed miRNAs in wheat**

591

592 **Table 1 Overview of functional modules in iwa-miRNA**

Module	Tool	Input	Output	Application
MiRNA Compilation	miRNARetrival	Name of species and miRNA databases	Already annotated miRNAs	Aggregate annotated miRNAs provided by four representative miRNA databases
	genomePrepare	Name of species or genome sequences and annotation	Path of formatted genome sequences and annotation	Get genome sequences and annotation
	miRNAPredict	SRA accession numbers or uploaded fastq files	Predicted miRNAs	Predict miRNAs from sRNA-Seq data
	miRNATranslate	Output from miRNARetrival and miRNAPredict	miRNA and miRNA precursors with a uniform format	Translate annotated and predicted miRNAs into the genomic coordinate system
MiRNA Selection	miRNASelection	Output from miRNATranslate	Selected miRNAs	Select promising miRNA candidates
Manual Curation	manualCuration	Output from MiRNA Selection	Summary and report pages	Determine the quality of selected miRNAs

593

594

595

596

597

598

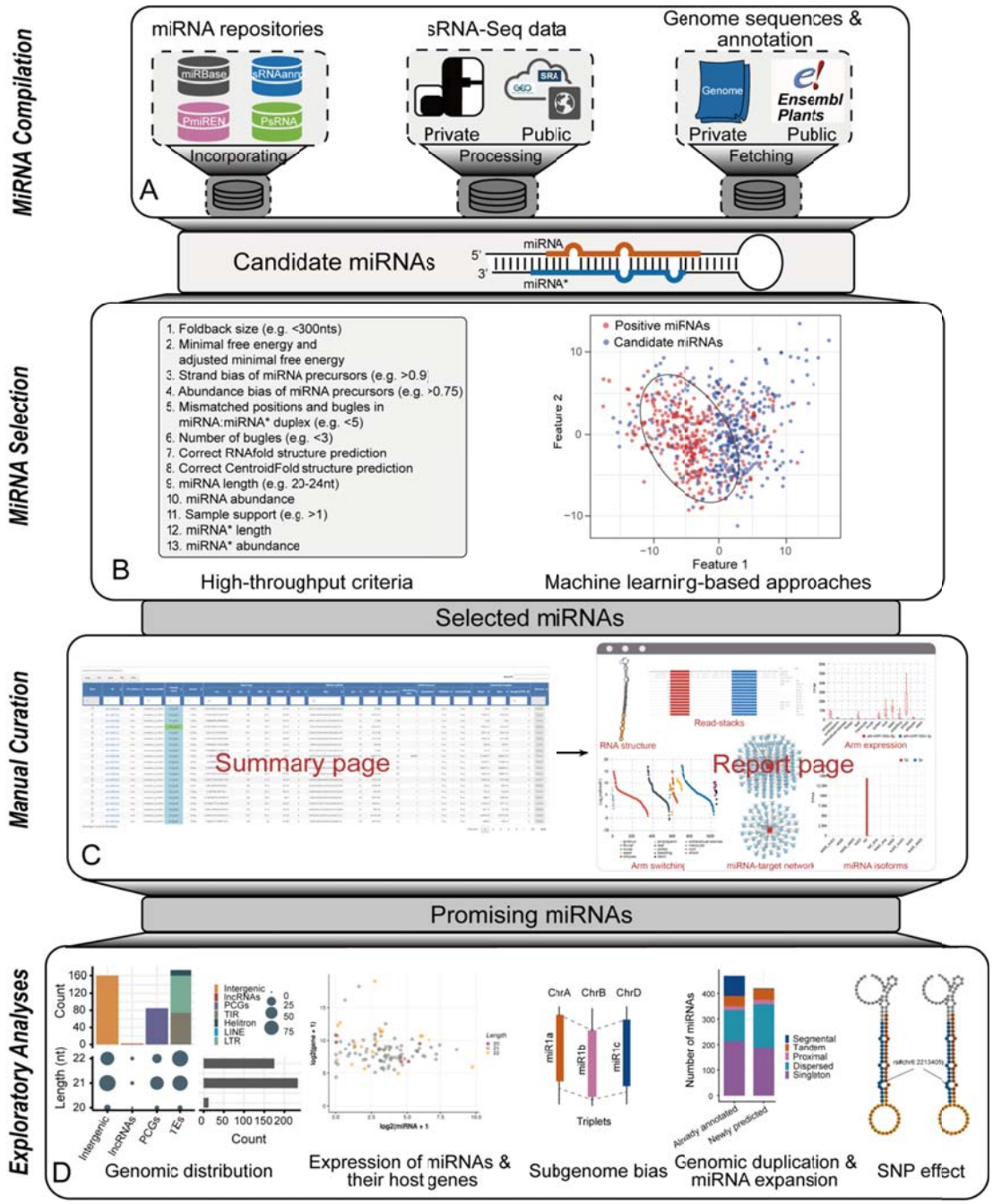

Figure 1 Graphical summary of iwa-miRNA

A. MiRNA candidates are generated by aggregating annotated and predicted miRNAs. B. Promising miRNAs are selected using high-throughput criteria and machine learning approaches. **C.** Manual curation of selected miRNAs based on annotation information from summary and report pages. **D.** Exploratory analysis of selected miRNAs. PCGs, protein coding genes; TEs, transposable elements; TIR, terminal inverted repeat; LINE, long interspersed element; LTR, long terminal repeat.

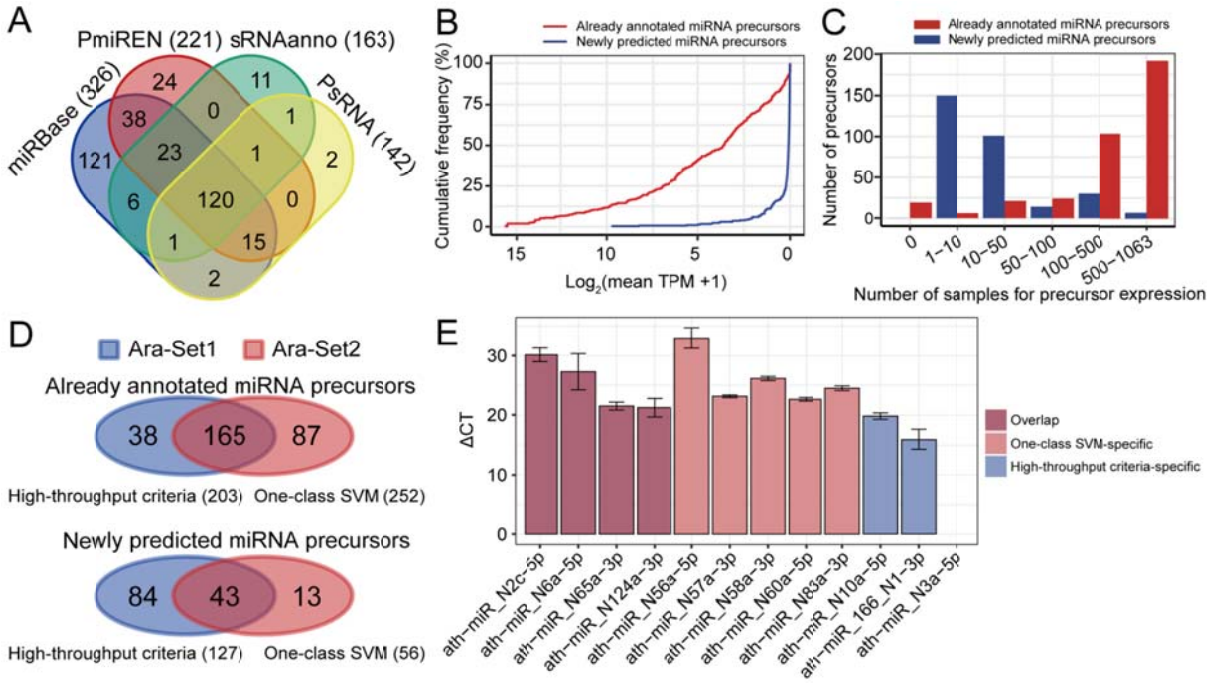

Figure 2 Application of miRNA annotation in *Arabidopsis*

A. Venn diagram comparing miRNA precursors provided by four miRNA databases. **B.** Cumulative frequency of log2 expression levels of already annotated and newly predicted miRNA precursor candidates. **C.** Number of expressed miRNA precursor candidates (already annotated and newly predicted) in different samples. **D.** Venn diagram comparing miRNA precursor candidates between the two miRNA selection approaches. **E.** Quantitative RT-PCR (qRT-PCR) results for 11 candidates. CT values were normalized to U6 small RNA. All values represent means ± SE (n = 3). TPM, transcripts per million; SVM, support vector machine; CT, cycle threshold.

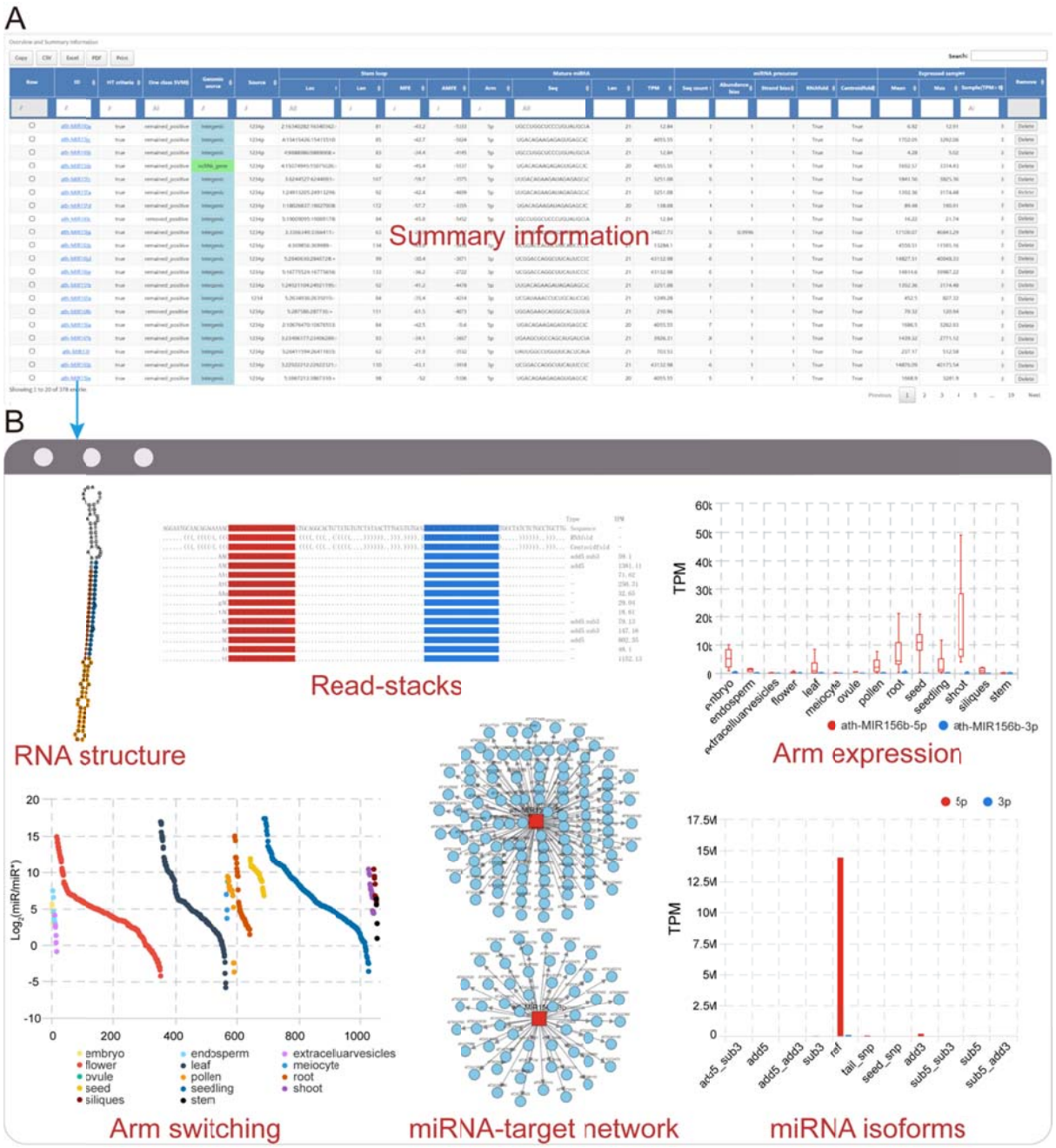

Figure 3 Summary and report pages generated by iwa-miRNA

A. Screenshot of summary page reporting the information of some features for *Arabidopsis* miRNAs. **B.** The report page for ath-MIR156b. HT, high-throughput; MFE, minimal free energy; AMFE, adjusted minimal free energy.

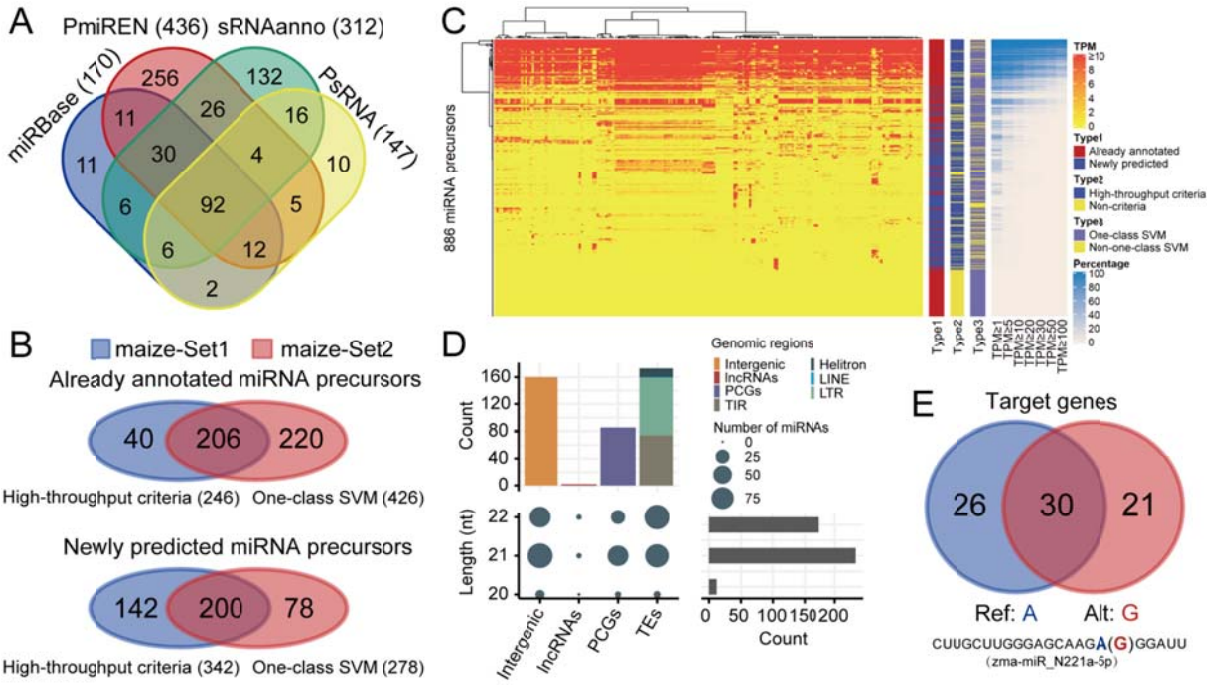

Figure 4 Application of miRNA annotation in maize

A. Venn diagram comparing miRNA precursors provided by four miRNA databases. **B.** Venn diagram comparing miRNA precursor candidates generated by two miRNA selection approaches. **C.** Expression levels of 886 miRNA precursors in 195 samples. Adjacent bar chart indicates three categorical results. The right panel shows the percentage of samples with expression greater than different thresholds (e.g. 1, 5, 10, etc.) **D.** Distribution of different miRNA lengths among different genomic features. **E.** The effect of SNP rs#chr6:22134053 on target genes of zma-MIR_N221a.

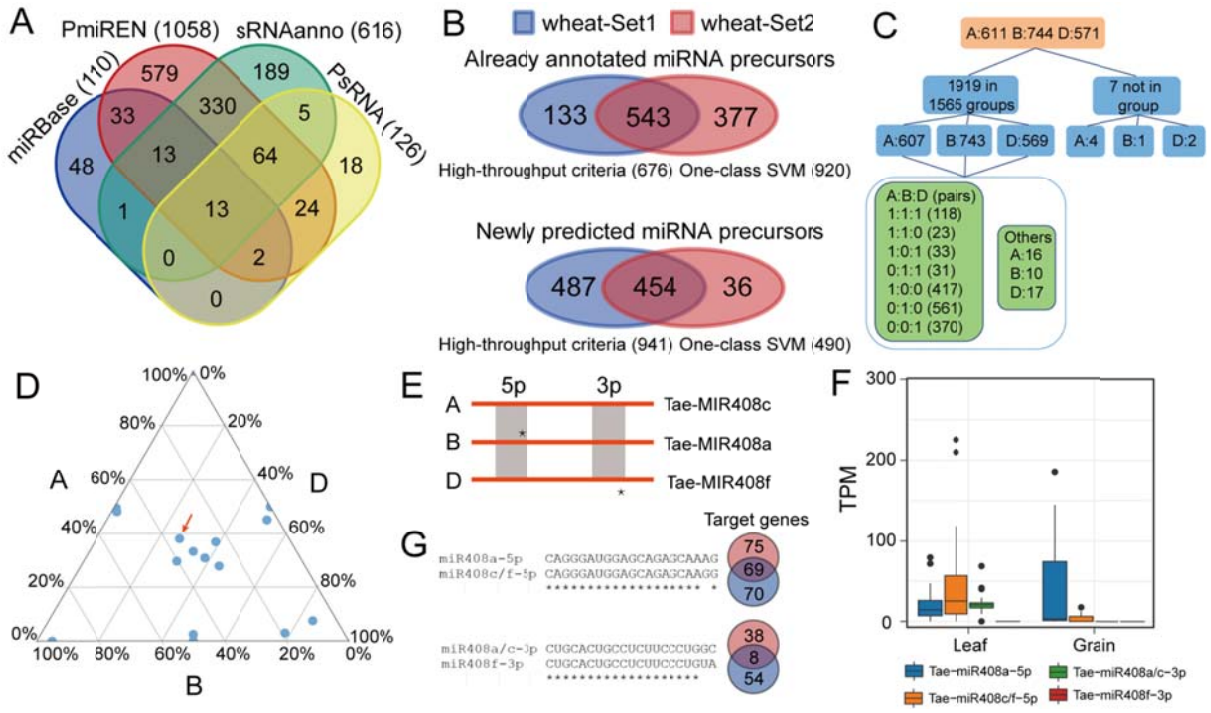

Figure 5 Application of miRNA annotation in wheat

A. Venn diagram comparing miRNA precursors provided by four miRNA databases. **B.** Venn diagram comparing miRNA precursor candidates generated by two miRNA selection approaches. **C.** Number of homologous groups and miRNAs in different A:B:D configurations. **D.** Ternary plot of miRNA expression levels from A, B, and D in triads. Tae-MIR408a/c/f is indicated by a red arrow. **E.** Schematic diagram of Tae-MIR408a/c/f in the ABD subgenomes. **F.** Expression levels of four mature sequences of Tae-MIR408a/c/f in leaf and grain tissues. **G.** The nucleotide difference and target alteration between mature sequences of Tae-MIR408a/c/f.