

Degradome Assisted Plant MicroRNA Prediction under Alternative Annotation Criteria

Salma Alzahrani, Christopher Applegate, David Swarbreck, Tamas Dalmay, Leighton Folkes and Vincent Moulton

Abstract— Current microRNA (miRNA) prediction methods are generally based on annotation criteria that tend to miss potential functional miRNAs. Recently, new miRNA annotation criteria have been proposed that could lead to improvements in miRNA prediction methods in plants. Here, we investigate the effect of the new criteria on miRNA prediction in *Arabidopsis thaliana* and present a new degradome assisted functional miRNA prediction approach. We investigated the effect by applying the new criteria, and a more permissive criteria on miRNA prediction using existing miRNA prediction tools. We also developed an approach to miRNA prediction that is assisted by the functional information extracted from the analysis of degradome sequencing. We demonstrate the improved performance of degradome assisted miRNA prediction compared to unassisted prediction and evaluate the approach using miRNA differential expression analysis. We observe how the miRNA predictions fit under the different criteria and show a potential novel miRNA that has been missed within *Arabidopsis thaliana*. Additionally, we introduce a freely available software ‘PAREfirst’ that employs the degradome assisted approach. The study shows that some miRNAs could be missed due to the stringency of the former annotation criteria, and combining a degradome assisted approach with more permissive miRNA criteria can expand confident miRNA predictions.

Index Terms— *Arabidopsis*, degradome, dicer, microRNA (miRNA) prediction, next-generation sequencing (NGS), parallel analysis of RNA ends (PARE), software.

1 INTRODUCTION

Small RNAs (sRNAs) are short, non-coding RNAs that have been found to regulate the expression of genes that are known to be involved in many diverse plant biological processes, from growth and development to environmental adaptation and stress response [1], [2], [3]. Their function takes place when they are loaded into an RNA induced silencing complex and guide it to silence their messenger RNA (mRNA) targets [4], [5]. In plants, it is typical for a high degree of complementarity between the sRNA and its target mRNA, often resulting in its translational silencing through cleavage and degradation [2], [6]. In recent years, next generation sequencing (NGS) technologies have been used to capture sRNA mediated cleavage fragments, often termed the degradome, using the Parallel Analysis of RNA Ends (PARE) protocol [7]. The PARE protocol captures the uncapped 5' ends of cleaved mRNA sequences providing

a snapshot of the mRNA degradation profile. The cleaved mRNA fragments within the profile can be aligned to reference transcripts, producing quantitative signals of varying strength that are position specific and indicative of sRNA mediated cleavage events. In addition, NGS has made it possible to capture an organism's entire sRNA profile in a single experiment on a genome-wide scale. Such a profile, often called the ‘sRNAome’, contains many classes of sRNA which are grouped based on their biogenesis and function.

MicroRNAs (miRNAs) are one such class of sRNA present within a profile and typically have a sequence length in the range of 20-24 nucleotides [8]. The defining feature of a miRNA is the precise excision of a double stranded region, sometimes called a ‘duplex’, from an RNA hairpin precursor structure by a Dicer like-enzyme [9], [10], [11]. The duplex contains both a mature miRNA sequence and a miRNA-star (miRNA*) sequence with a 2 nucleotide (nt) overhang at the 3' ends. However, a pairing between a mature miRNA and a miRNA* within a duplex is often imperfect, including variation in the number of nt mismatches, bulges, and the number of nt within a bulge. The full-length precursor of a miRNA also exhibits variation in features such as its stem-loop folding composition as well as its length [12]. Such variability of features within both a miRNA stem-loop precursor and a mature miRNA duplex in plants can present a challenge for the accurate computational annotation of miRNAs within a genome-wide profile and in particular the correct attribution of sRNAs to the

- S. Alzahrani and C. Applegate are with the School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom. Email: s.alzahrani@uea.ac.uk, chris.applegate.academic@gmail.com.
- D. Swarbreck is with the Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, United Kingdom. Email: David.swarbreck@earlham.ac.uk.
- T. Dalmay and L. Folkes are with the School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom. Email: t.dalmay@uea.ac.uk, l.folkes@uea.ac.uk.
- V. Moulton is with the School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom. Email: v.moulton@uea.ac.uk.

(Correspondence authors: Leighton Folkes; Vincent Moulton.)

class of miRNA [13].

The identification of miRNAs has been a subject undergoing intense study for the last decade [14]–[16]. Many of these investigations involve NGS sRNA data and a reference genome that allows researchers to apply computational methods to identify miRNAs and their function on a genome-wide scale [17], [18]. As a result, several miRNA repositories were developed to archive miRNA biogenesis and functional information such as miRBase [19] and PmiREN [20]. The methods used within most tools for predicting plant miRNAs, such as miRPlant [21], miRCat2 [22], and miRDeep-P [23], were designed using a set of suggested miRNA biogenesis features. These features comprised a set of stringent criteria that have been used to model a miRNA. However, these criteria were published over a decade ago [13] and do not describe, or account for, a growing number of validated miRNAs that follow a model composed of a less stringent criteria. Therefore, the currently available tools based on an outdated criteria risk discarding *bona fide* miRNAs. In spite of the use of a stringent biogenesis model underpinning their prediction algorithms, some of these tools still tend to generate a large number of false positive predictions [24].

Recently, a new set of miRNA annotation criteria has been reported [18], suggesting that more flexibility is required in some of the criterion of the standard miRNA annotation model. The newly suggested model also applies some restrictions on the length of the miRNA, miRNA* and precursor, and requires biological replication of the sRNA profile and suggests that further experiential validation beyond high-throughput sequencing of the sRNA profile is not required. In addition, the authors have suggested that these restrictions in their updates could contribute to the reduction of false positives. Even so, a more flexible choice of a less stringent set of parameters, e.g. allowing more mismatches within a duplex and increasing the size of gaps within a duplex, is likely to result in increasing the total number of miRNA predictions overall.

In this study, we revisit the sRNAome under the revised annotation criteria by exploring the effect of applying the newly suggested criteria on miRNA prediction using existing miRNA prediction tools. We focus on *Arabidopsis thaliana* due to its easily accessible resources: The *A. thaliana* genome is well described in The Arabidopsis Information Resources (TAIR) [25], and well-studied in terms of published miRNA annotations. We shall investigate the effect of using more permissive parameters for miRNA prediction, and in particular, we develop an algorithm to explore and evaluate different parameter combinations. Moreover, we present a new combination approach to miRNA prediction, which uses the functional information extracted from a genome-wide degradome-assisted sRNA target analysis.

Our new approach using degradome data helps in the sRNA annotation effort in several ways. Firstly, by conceptually reducing the number of sRNA candidates to those that are potentially functional and cleavage capable. Secondly, the use of less-stringent miRNA secondary structure prediction parameters for miRNA candidates within the functional sRNA subset becomes feasible when modulating by their function. And thirdly, the predicted miRNA

mediated cleavage signal and biogenesis information can be examined simultaneously to derive a final consensus miRNA candidate set that can be computationally filtered and ranked by confidence information for further experimental validation. We demonstrate that even though a greater number of candidate miRNAs tend to be generated with more flexible parameters, our combination method is able to reduce this number by employing degradome information. Our combination approach is made freely available in user-friendly software called ‘PAREfirst’ that can be downloaded from: <http://srna-workbench.cmp.uea.ac.uk>, and the open source code is available at: https://github.com/sRNAworkbench/UEA_sRNA_Workbench/.

2 MATERIALS AND METHODS

2.1 Data sets

We used three wild-type and three Dicer-like1 (DCL1) mutant *A. thaliana* sRNA biological replicates that are publicly available (GSE90771) [22], we called the wild-type sRNAome WTA, WTB, and WTC, and the DCL1 mutant samples: DCL1A, DCL1B, and DCL1C. The PARE analysis was performed using the corresponding degradome for each wild-type replicate that are also available on GEO (GSE113958) [26]. For evaluation, we used the 326 unique mature miRNAs, which are excised from 426 precursors for *A. thaliana* from the miRBase registry v22, and for the sake of this paper, we refer to them as the validated miRNAs. The reference *A. thaliana* genome TAIR10 [25] was used, in addition to the transcriptome containing the cDNA for the updated representative gene model [27].

2.2 Methods

We begin by describing our approach to investigate the effect of allowing more permissive parameters on miRNA prediction. Our approach uses the miRCat2 tool for miRNA prediction and the PAREsnip2 [26] tool for degradome analysis. In brief, miRCat2 is a miRNA prediction method that uses an entropy-based approach to detect miRNAs within a genome. As inputs, the method requires a reference genome and sRNAome. The method first identifies potential miRNA candidates based on sRNA abundance and then applies a number of filters such as mapping locus, size class distribution and miRNA-like alignment patterns on the candidates before calculating their miRNA secondary structures. The method outputs miRNA predictions in a tabular format and was selected for its improved accuracy when compared to similar tools. PAREsnip2 is a degradome analysis method that can be used to identify sRNA targets. As inputs, the method requires a degradome, sRNAome and a set of transcripts often called a ‘transcriptome’. The method first performs several optional quality filtering steps on the input sequences. The method’s algorithm then encodes input sequences into a decimal number which is then used to make exact match sequence alignments and subsequently identify potential sRNA-target pairs. The tabular output contains information on the sRNAs and their potential target sites along with abundance and degradome assisted confidence

metrics. PAREsnp2 was selected because of its configurability and its ability to process large volumes of data without imposing substantial computational time and resource constraints. Both of these tools are implemented in the UEA sRNA workbench, and have the advantage that they can be easily configured and have been shown to perform comparatively well compared with other tools [28]. First, we implemented a parameter-search algorithm that is described in detail in the next section to produce a collection of roughly 150 exploratory parameter sets (denoted EPS) that are listed in Supplementary Table S1, which can be found on the Computer Society Digital Library. We then produced an updated miRCat2 parameter set (denoted UPS) based on the new criteria presented in [18]. For comparison, we present the main criteria for EPS and UPS in Table 1, available online, together with the default miRCat2 parameter set (denoted DPS). Then, for each wild-type sRNA sample we obtained three sets of miRNA predictions using the miRCat2 tool with the DPS, UPS, and EPS.

Next, we performed a target analysis with PAREsnp2 and used its outcome to control the false positive miRNA predictions that could result from relaxing the biogenesis parameters without losing the majority of the validated miRNAs. The analysis was performed using the wild-type sRNA replicates, degradome replicates, the transcriptome, and the genome. In addition, we used Fahlgren and Carington [29] targeting rules, allowed categories 0-3, disabled MFE and p-value filters, sRNA length from 18-25 nt, and disabled the core region multiplier. These more permissive parameters were used to capture validated sRNA-target interactions that would have been missed using the default settings [30].

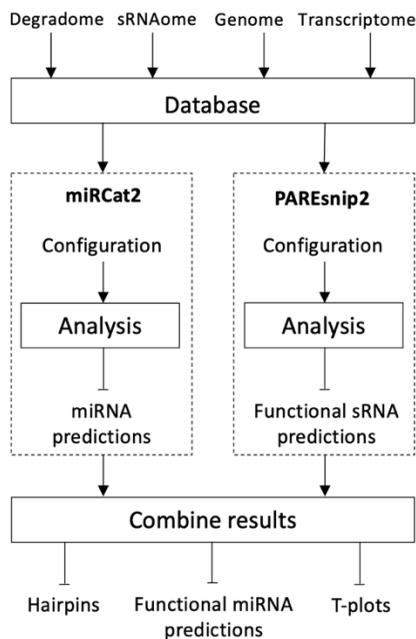


Fig. 1. Schematic of the PAREfirst workflow used to perform a large-scale investigation of miRNAs and their targets evidenced through the degradome. Solid rectangles represent processes, arrowed lines represent inputs and data flow, and lines represent output. The modules within PAREfirst are enclosed within dotted lines.

TABLE 1
THE miRCat2 PARAMETERS FOR DPS, UPS, AND THE RANGES FOR EPS PARAMETERS

miRCat2 parameter	DPS value	UPS value	EPS ranges
min_length (a)	20	20	18, 19, 20
max_length (b)	23	24	23, 24, 25, 26
min_fold_len (c)	45	40	40, 45
max_fold_len (d)	250	300	250, 300, ..., 400
max_amfe (e)	-22	-22	-32, -27, ..., -2
Complex (f)	0.90	0.90	0.50, 0.60, ..., 0.90
clear_cut_perc (g)	0.92	0.90	0.52, 0.62, ..., 0.92
gaps_mirna (h)	4	5	4, 5, ..., 8
no_loop (i)	3	3	3, 4, ..., 7
Repeats (j)	25	25	25, 30, ..., 40
p-val (k)	0.05	0.05	0.05
RANDfold (l)	false	false	false
complex_loop (m)	true	true	true

Parameters are labelled as follows: (a) minimum length of miRNA, (b) the maximum length of miRNA, (c) minimum length of precursor, (d) maximum length of precursor, (e) maximum value for the adjusted MFE for a miRNA precursor, (f) complexity of sequence, (g) percent of incident reads that should fall between the same start and end positions as the miRNA, (h) maximum number of consecutive gaps on the precursor on the miRNA location, (i) Maximum number of bulges in the loop area of the precursor, (j) maximum number of times a sRNA can map to a genome, (k) RANDfold computation, (l) threshold for the RANDfold value, and (m) if a precursor with multiple loops between miRNA and miRNA* is allowed.

Moreover, we used the DCL1-mutant sRNA data sets to validate the predicted functional miRNAs. In particular, we performed a differential expression (DE) analysis using DESeq2 [31] within iDEP9 [32] between the three wild-type sRNA replicates and the three DCL1-mutant sRNA replicates. The predicted miRNAs that were two-fold down-regulated in the DCL1-mutant were considered as enriched candidate miRNAs. Additionally, we performed a similar DE analysis between the wild-type replicates and DCL4-mutant triplicates that were obtained from GEO (GSM4061704, GSM4061705 and GSM4061706) [33]. We show the results of DE analysis for WT_DCL1 in Supplementary Table S2, and for WT_DCL4 in Supplementary Table S3, available online.

To further investigate the enriched miRNA candidates, we aligned the candidates to all the plant species miRNAs that were retrieved from miRBase using PatMaN [34], allowing one mismatch to allow for isomiRs. We also discarded the candidates that align to other RNA classes such as transfer-RNAs, ribosomal-RNAs, small-nuclear-RNAs and small-nucleolar-RNAs.

2.3 Generating permissive miRCat2 parameter sets

To generate exploratory parameter sets (EPS), we designed an iterative local parameter search method that uses a hill-climbing algorithm [35] to explore more permissive EPS combinations. More specifically, ranges were set for each configurable miRCat2 parameter (Table 1), which were provided to the algorithm. The algorithm starts with a random selection of parameters within the given ranges. For

each iteration, a check is made to each neighbouring EPS in which each parameter value is incremented or decremented. Predictions using miRCat2 were then made for each of the neighbouring EPS and the algorithm chooses a new EPS based on the score function described below. This process is repeated until there is no further improvement on the score and the highest scoring EPS is retained. The algorithm was performed 100 times, using a randomly selected starting EPS for each run, of which we selected the highest scoring 50 EPS. The search was applied on each wild-type sRNA replicate, and we combined the top scoring 50 EPS to generate our final collection of 150 EPS that are listed in Table S1.

The score of each EPS was calculated based on three sets: the set of miRCat2 predictions using the EPS being evaluated denoted by m , the set of predicted functional sRNAs from PAREsnip2 denoted by p , and the set of validated miRNAs from miRBase denoted by mb . For the purpose of the parameter search method, we obtained the set p for each replicate from PAREsnip2 using Allen targeting rules [36] and the default parameters with the exclusion of weak cleavage signals summarised as PAREsnip2 categories 2, 3 and 4 interactions to generate high confidence results. The score that we used is given by:

$$\text{score}(\text{EPS}) = a + b + (a / (a + b) * 1.5) + (c / (c + d)), \quad (1)$$

where $a = |m \cap p \cap mb|$, $b = |m \cap p \cap mb'|$, $c = |m \cap p' \cap mb|$, and $d = |m \cap p' \cap mb'|$, where $'$ denotes the set complement.

In particular, the score in (1) was calculated based on the intersection between the predictions of m and p , in addition to the ratio of mb in that intersection multiplied by 1.5 as we sought EPS with an improved number of validated functional miRNA predictions. We also considered the non-functional validated predictions in m , hence, we added the ratio of non-functional predicted mb to the total non-functional predictions in m .

2.4 PAREfirst implementation

We designed and implemented a method called PAREfirst (Fig. 1), which combines miRCat2 and PAREsnip2 into a tool within the UEA sRNA Workbench now mentioned above. The user can perform a highly configurable analysis in a user-friendly interface that produces an easily interpreted list of predicted miRNAs along with a visual representation of the prediction secondary structure and confidence metrics. PAREfirst accepts as an input the following files: sRNAome, degradome, transcriptome, and a genome in a FASTA format. Also, the user can configure the parameters for both PAREsnip2 and miRCat2 before starting the analysis. The tool performs the PARE analysis first using PAREsnip2 to produce the functional sRNAs. After that, it generates the miRNA predictions with miRCat2. The tool then combines the results to gain a set of functional miRNA predictions. Importantly, this method allows the use of less stringent rules for miRCat2, since the outcome of miRCat2 is controlled by the PARE analysis results. As an output, PAREfirst exports and displays a table within a Graphical User interface containing information for the predicted functional miRNAs. In addition, the user is able

to visualise and export target-plots (t-plots) [7] that are useful to distinguish true target sites from background noise, and the secondary structures for the predicted miRNA precursor using RNAplot [37] to visualise the hairpins. To assess the performance of PAREfirst, we ran it on our data sets and benchmarked the results against other commonly used miRNA detection tools, miRCat2 and miRDeep-P2 [38]. To produce PAREfirst predictions, we used the EPS for miRCat2 analysis and the permissive PAREsnip2 parameters that were described above. The miRNA predictions for each of the other tools were obtained using the UPS for miRCat2 and the default miRDeep-P2 parameters (described in the user manual) for miRDeep-P2. We aligned the predictions to the mature miRNAs from miRBase to identify the validated miRNAs. Also, we excluded the miRNA candidates that aligned to other RNA classes.

3 RESULTS

To compare the effect of applying alternative annotation criteria to miRNA prediction, we ran miRCat2 on wild-type sRNA data sets: WTA, WTB, and WTC, using the three parameter sets: DPS, UPS and EPS. For the purpose of improved confidence in predicted miRNAs, those having fewer than 10 reads were discarded from further analysis. We also filtered read counts by excluding isomiRs (sequences that are one or two nt shorter or longer than the canonical mature miRNAs), thus providing a clear quantification of the mature miRNA for each prediction. In addition, we considered conservation of the mature miRNA sequence across the three biological replicates in an attempt to provide a higher degree of confidence based upon multiple observation of the sequence [39]. For this study, we define a miRNA as conserved if it was expressed in at least two out of the three wild-type replicates. Furthermore, we used the validated miRNAs from miRBase as a reference to evaluate the results, even though we acknowledge its limitations with regards to the quality of miRNA annotations [40], [18]. We split the candidates into two groups: C and P; the reason for this categorization is that multiple miRNAs and miRNA-like RNAs can originate from one miRNA precursor [41], [42]. Moreover, we present the number of miRNAs that are shared between the miRCat2 predictions using DPS, UPS and EPS in Supplementary Fig. S1, available online.

We present the number of miRCat2 predictions using the three parameter sets in Table 2a. The table includes the number of all validated miRNAs within a sRNA replicate as well as the validated miRNAs predicted by miRCat2 using the three parameter sets. Comparing the results produced using each of the parameter sets, we observe that the UPS succeeded in predicting one or two more of the validated miRNAs in each replicate, and only predicted several new candidates when compared to DPS. Hence, it is likely that the false positive rate is still low and the performance of UPS is overall sufficient. Yet, using the EPS with miRCat2 performed slightly better in predicting more of the validated miRNAs. In particular, four more validated miRNAs that are conserved were predicted compared to DPS. However, it generated a high number of new candidate predictions that may include a number of false

TABLE 2

THE NUMBER OF VALIDATED miRNAs THAT WERE FOUND IN OUR sRNA DATA SETS, miRcat2 PREDICTIONS USING DEFAULT PARAMETERS (DPS), UPDATED PARAMETERS (UPS), AND EXPLORATORY PARAMETERS (EPS) FROM THE PARAMETER-SEARCH METHOD.

Filter	Replicate	AV	DPS			UPS			EPS		
			V	C	P	V	C	P	V	C	P
(a) None (all predictions)	WTA	132	87	8	33	89	29	33	92	129	43
	WTB	126	85	6	28	87	25	31	89	97	39
	WTC	127	90	6	26	91	27	28	95	106	37
	Conserved	136	87	6	31	88	28	33	91	103	42
(b) PAREsnip2 filter	WTA	127	85	5	31	87	13	31	90	61	40
	WTB	109	74	4	25	75	9	28	77	35	33
	WTC	83	55	3	18	55	4	19	59	34	26
	Conserved	121	78	3	29	78	9	31	81	39	38
(c) PAREsnip2 and DCL1 filters	WTA	91	63	3	27	64	8	26	67	33	34
	WTB	78	55	2	22	56	6	24	58	26	29
	WTC	59	41	2	18	41	3	19	45	17	26
	Conserved	85	58	1	26	58	6	27	61	23	34
(d) Only DCL1 filter	WTA	95	65	5	29	66	17	28	69	77	37
	WTB	92	63	3	25	65	16	27	67	57	34
	WTC	93	67	3	25	68	17	26	72	57	34
	Conserved	97	65	3	28	66	18	29	69	63	37

AV: : all validated miRNAs within a sRNA replicate, V: validated miRNAs predicted by miRcat2, C: candidate miRNAs predicted by miRcat2 and do not map to miRBase validated precursor loci, P: candidate miRNAs predicted by miRcat2 that map to a validated precursor but do not map to the canonical miRNA site. The conservation level used is between two or three replicates. All validated and candidate miRNAs have a read count above 10 reads.

positives.

We now present the results of applying the PAREsnip2 filter on the miRcat2 predictions. Although using less-stringent parameters for predicting miRNAs and their precursors can introduce an increase in false positive predictions, we only consider the intersection between PAREsnip2 and miRcat2 predictions. Table 2b presents the number of miRNA predictions that are involved in a mRNA targeting interaction. It appears to be that the PAREsnip2 filter kept a similar number of the validated miRNAs in each of the miRcat2 results across the three parameter sets, as true miRNAs are more likely to have a target. On the other hand, it discarded one-third or more of the non-validated candidates predicted by DPS and UPS. The functional filter reduced the majority of the EPS candidates, and, upon manual inspection, we found that these candidates had secondary structures that were far beyond our reasonable expectation of a miRNA precursor, hence we consider this group to contain the highest number of potential false positive miRNA candidates. Accordingly, a function-first approach using degradome-assisted functional-filtering shows promising results, where we reduced the miRNA candidates to a list of 39 potential conserved functional miRNAs that can be carried forward for further investigation.

To validate the functional miRNA predictions, we performed a DE analysis between the wild-type and DCL1-mutant samples. Since the DCL1 has an important role in the miRNA biogenesis pathway in *A. thaliana*, knocking down its activity causes reduction in the expression of the miRNAs [43]. The outcome of applying the DCL1 validation filter was a set of predicted functional miRNAs that are enriched in the wild-type samples. The validation step discarded a number of the functional miRNAs and the results are shown in Table 2c. The outcome shows a further reduction in the candidate predictions, where these remaining candidates could have a higher degree of confidence. Full details for all functional miRNA candidates and

the conserved enriched functional miRNA candidates are found in Supplementary Tables S4 and S5, available online.

We also present the results for applying the DCL1 validation on all predictions in Table 2d. It seems that several validated miRNAs were not down-regulated in the mutant samples. This could be because of a DCL1-independent pathway, and, in some cases in Arabidopsis, miRNAs are sometimes processed by a different Dicer family member such as DCL4 [44]. To investigate this hypothesis, we applied the DE analysis between wild-type and DCL4-mutants. The outcome of applying the DCL4 validation filter on the predicted functional miRNAs are presented in Supplementary Table S6, available online. The comparison between the up- and the down-regulation that occurred in WT_DCL1 and in WT_DCL4 is shown in Supplementary Fig. S2, available online, which shows an overlap of seven validated miRNAs that were Down_WT_DCL1 and Up_WT_DCL4, and this could indicate a major involvement of DCL4 in some miRNA biogenesis pathways.

We further investigated the applicability of our method on other plants. Here we followed a similar approach for the data preparation and the target analysis parameters, however, we excluded the differential expression analysis due to the lack of DCL1-mutant data. The investigated species were the commonly studied Tomato, *Solanum lycopersicum* (*S. lycopersicum*), and Rice, *Oryza sativa* (*O. sativa*). For *S. lycopersicum* analysis, we used the publicly available sRNA data sets from GEO [45] (leaf GSM803579, flower GSM803580, and fruit GSM803581), and performed the target analysis with the corresponding tissue degradome data from a different study [46] (leaf GSM553688, flower GSM553689, and fruit GSM553690). The reference genome (SL3.0) and transcriptome (ITAG3.0) were downloaded from the Sol Genomics Network [47]. For *O. sativa* data, we used four sRNAome libraries [48] (Indica rice seedling and panicle, GSM562942 and GSM562943, Japonica rice seedling and panicle, GSM562946 and GSM562947). Also, we used two degradome libraries from another study [49]

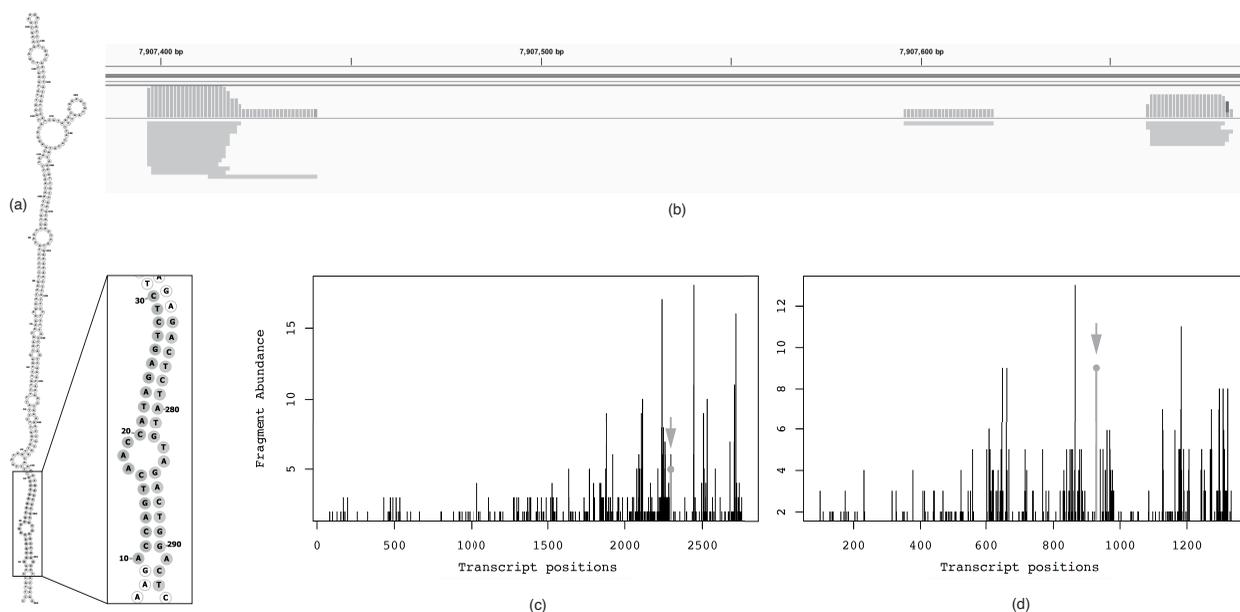


Fig. 2. The novel miRNA candidate hairpin precursor with its mature miRNA and miRNA* sequences. The coordinates of the precursor locus within *A. thaliana* genome, Chromosome4 is 7907388-7907687(+). (a) secondary structure of the precursor where the miRNA arises from the left arm of the hairpin and the miRNA* rises from the right arm [52], (b) a coverage plot of the precursor locus where miRNA alignments are presented on the left side and miRNA* alignments are presented on the right side of the plot, (c) t-plot of mature miRNA target mRNA AT4G00340.1, and (d) t-plot of miRNA* target mRNA AT3G61790.1 where arrows point at the cleavage positions.

(seedling GSM455938 and panicle GSM455938), where we performed the target analysis on one tissue of the degradome data with the two corresponding tissue sRNA data sets. The reference genome and transcriptome were obtained from the Rice Annotation Project Database [50], [51]. The annotation for both *S. lycopersicum* and *O. sativa* were performed using the mature miRNAs from the miR-Base registry v22. As for *S. lycopersicum* results, the numbers of found mature miRNAs within the data sets were low, this explains the low numbers of miRcat2 predicted validated miRNAs. However, it appears that the target analysis kept the validated miRNAs and excluded the majority of candidate miRNAs. In *O. sativa*, there was a noticeable difference between the replicates result. Overall, there is a slight increment in the number of predicted validated miRNAs in the UPS results compared to the DPS. The functional filter also excluded a minority of the predicted validated miRNAs and excluded a majority of the candidates. Correspondingly, these results suggest that our method is broadly applicable to other plant species. A summary of the analysis results for *S. lycopersicum* and *O. sativa* are presented in Supplementary Table S7, available online.

3.2 Investigation of the miRNA annotation criteria

We investigated whether the predicted secondary structures produced by miRcat2 with DPS, UPS, and EPS fit under the 2008 criteria, 2018 criteria, both criteria, or do not fit any of them, where the category of both criteria does not intersect with the 2008 or the 2018 criteria. The results are presented in Supplementary Table S8, available online. In this table, we refer to the category where the hairpin does not fit any criteria as 'Undefined'. In the following, we only consider miRNA hairpin rather than the unique mature miRNA sequences, since the annotation is based on both the miRNA/miRNA* duplex and hairpin structure

features. We observed that most of the miRNAs that were predicted exclusively from any single replicate do not fit any criteria, hence, they were discarded since the conservation between replicates supports the confidence of the miRNA. Therefore, Supplementary Table S8, available online, only includes the grouping of the precursors that were predicted by miRcat2 in two or three sRNA replicates.

We observed that a few validated miRNAs do not fit any of the criteria, and a similar case was addressed by Axtell and Meyers [18], where some entries in miRBase may need to be revised. In addition, we looked into the validated miRNA hairpins that are shown in miRBase, and we observed that some of them have 1-nt overhangs at the 3' ends of the miRNA/miRNA* duplex instead of 2-nt as the annotation criteria suggested. Interestingly, even with the 2018 criteria included via the UPS for miRcat2, some of the predictions are still categorised as 'Undefined' criteria.

As presented previously, Supplementary Table S8 also shows that the EPS predicts a higher number of candidate precursors than the DPS and UPS, and it seems that the majority of these candidates fit under the 'Undefined' category, and these candidates were considered to contain a high number of the false positives. Using the degradome-assisted sRNA targets as a filter has discarded the majority of these false positives, while keeping most of the validated miRNA precursors. Furthermore, the majority of DPS predictions were kept during this step, where these predictions provide confidence through their strict miRNA features.

The next step of validation is the differential expression analysis between the wild-type and DCL1-mutant samples, where we used the miRNAs that are enriched in the wild-type samples as a second layer to filter the miRNA predictions. As resulted from the PAREsnip2 filter, the

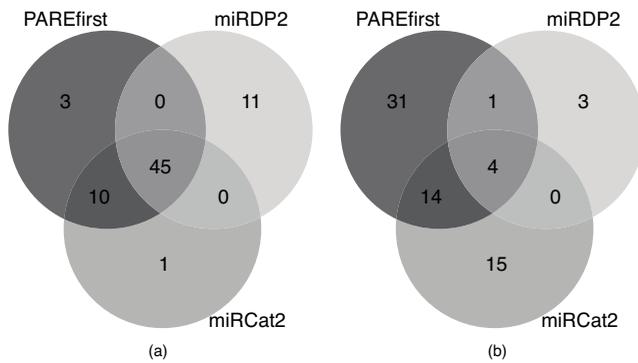


Fig. 3. Venn diagram showing the intersection between the number of (a) validated miRNAs, and (b) candidate miRNAs that were predicted by PAREfirst, miRDeep-P2 (miRDP2), and miRCat2 in the three sRNA replicates.

DCL1 validation also excluded the majority of the EPS candidate miRNAs that fit under the ‘Undefined’ category, and only excluded a minority from the rest of the results.

3.3 New miRNA and miRNA* candidates

We carried out an investigation of the miRNA candidates that are involved in a mRNA-target interaction and enriched in the wild-type samples. Before doing so, we excluded miRNAs that map to mature miRNAs, mature miRNA isomiRs, and the other RNA classes. As a result, we identified a potential novel miRNA with its miRNA* that map to one unannotated locus in the genome. We performed a local alignment on all the plant species miRNAs from miRBase, and we found that the miRNA and miRNA* are not present in other genomes. The mature candidate was predicted along with its miRNA* and hairpin structure using UPS and EPS for miRCat2. The secondary structure of the candidate precursor with the highlighted miRNA/miRNA* duplex is shown in Fig. 2a. In addition, the majority of reads in the precursor align to the miRNA/miRNA* duplex as shown in Fig. 2b. This candidate precursor falls under the 2018 miRNA annotation criteria. The abundance of the mature miRNA appears to be low (less than 100 reads) compared to most of the known miRNAs in our samples. Additionally, the mature abundance is double the miRNA*, which is a requirement for a well formed ‘bona fide’ miRNA duplex [18]. As shown in the WT_DCL1 DE analysis results in Supplementary Table S2, available online, both the mature and the star sequence are differentially expressed with $\text{Log}_2(\text{fold-change})$ of 4.48 and adjusted p-value < 0.05 . We present the t-plots for the most confident predicted mRNA-target interactions for the mature miRNA in Fig. 2c and the miRNA* in Fig. 2d. It seems that the mature sequence does not show a strong signal in its t-plot compared to the miRNA*, which shows a higher peak. According to PAREsnip2 results, the mature miRNA is predicted to be involved in targeting interactions of category 2 with three different genes, while the miRNA* have six different targeting interactions of category 2. Details about the candidate precursor and the miRNA/miRNA* target interactions are found in Supplementary Table S5, available online. Additionally, we

explored alternative four *A. thaliana* data sets (flower GSM707678, leaf GSM707679, root GSM707680, and seedling GSM707681), and the potential novel miRNA and its star sequence were present in flower, leaf, and seedling samples. We performed miRCat2 analysis on these three data sets using UPS and the potential novel miRNA secondary structure was predicted with its miRNA/miRNA* duplex. The miRCat2 results for these new samples are presented in Supplementary Table S9, available online.

Furthermore, there are a number of miRNA candidates that derived from validated precursors in miRBase. These miRNAs are predicted with their complementary miRNA* (Supplementary Table S10, available online), however, these miRNA* sequences are not registered in miRBase. Moreover, there are other studies that acknowledged these miRNA* [53], [54].

3.4 PAREfirst comparison

To investigate how PAREfirst compares to the existing traditional miRNA prediction tools, we ran our data sets through PAREfirst, miRCat2 and miRDeep-P2. We present the output of each tool in Supplementary Tables S11, S12 and S13, respectively. Fig. 3 shows the overlap of the number of validated and candidate predictions among the three tools, where these predictions are conserved between two or three out of three replicates. Fig. 3a indicates that PAREfirst captures the majority of the other tool’s validated miRNA predictions, in addition to other three miRNAs that were not identified using the updated miRNA annotation criteria. Fig. 3b also shows that PAREfirst allows the identification of candidate miRNAs that as observed in this study, do not necessarily conform with the standard model.

4 CONCLUSION AND DISCUSSION

Axtell and Meyers [18] argued that a change in miRNA annotation criteria is necessary since some validated miRNAs are missed by the former criteria and some novel miRNAs might be missed too. Hence, there is a need to update the parameters used within the existing prediction tools, or design new tools that incorporate less stringent rules. By systematically exploring different parameter sets, it is clear that flexible parameters have an impact on miRNA prediction, and we should keep the balance between predicting miRNAs with novel features and an overestimation of the miRNA profile. We showed that using the updated and the less stringent criteria increases the capture of validated miRNAs while keeping a similar number of potential candidates through filtering with the degradome analysis. We applied our method to several publicly available *A. thaliana* sRNA data sets and we were able to identify a potentially novel miRNA candidate that has been previously missed by tools that are dependent on outdated miRNA-rule sets.

A. thaliana is a well-studied genome, and we can expect that the miRNA profile is well characterised, yet, we have been able to identify a potential novel miRNA. Applying our approach on less well annotated plant genomes could capture not only the miRNAs that are easily identified through existing methods but also the miRNAs that would

otherwise be missed due to their extreme biogenesis characteristics. With this in mind, the variance within the parameters that identify novel miRNAs in *Arabidopsis* may not be the most suitable in all cases, and we hypothesise that improvements in annotation results could be obtained from investigating species-specific parameter sets. To this end, we have provided some software to enable researchers to take this forward in the model species of their choice.

The degradome analysis is a high-throughput approach to identify miRNA mediated cleavage [55], and the defined category system within degradome analysis tools ranks the confidence level of miRNA-mRNA interactions. A degradome sequencing experiment offers many advantages when compared to low-throughput methods typically used for miRNA target validation, such as 5' RACE (rapid amplification of cDNA ends) [56]. The advantages are not only present in time-cost savings, but also the global nature of the degradome profile being captured in a single experiment can reveal multiple miRNA-target interactions, useful for building miRNA-mediated gene regulatory networks [57], [58]. In addition, when compared to traditional sequence similarity approaches for target prediction [59], using a degradome assisted miRNA target prediction approach provides valuable quantitative confidence values based on experimental evidence [26], [39], [60], [61]. Our results show that most of the enriched known and candidate miRNAs are predicted to be functional, which suggests that the degradome analysis provides useful supporting evidence for identifying functional miRNA candidates without using further validation steps. Be that as it may, a degradome assisted approach is somewhat limited by its dependence on the expression and tandem capture of the miRNA and miRNA mediated cleavage signal within the sRNAome and degradome data sets. The expression of many miRNAs and their targets are localized both temporally and spatially, specific to factors such as tissue, growth, and environment. However, testing for condition specific miRNA candidates and their targets is a common goal in sequencing experiments that investigate within and for such factors [62]–[65]. Also, generating the degradome data required by our method can be challenging and is not necessarily straightforward [66]. However, new and optimized degradome protocols are regularly becoming available [66]–[68]. As NGS techniques become more accessible, we envisage more degradome libraries will become available, enabling the use of our approach with more varieties of species.

The identification of biogenesis and function of a miRNA is important to understand its role within biological pathways and networks. Researchers and miRNA databases, such as miRBase, are enabling the provision of a reliable set of functional information that will enhance the advancement of the microRNA research field. In particular, miRBase is being improved not only by providing miRNA annotation entries, but by also including the functional information of these miRNAs [69]. Our functional approach moves toward the aim of identifying miRNAs and their target mRNAs, and we hope that it will have an impact on enriching the literature of miRNA functions.

In conclusion, our degradome-assisted method for

miRNA prediction appears to provide broader predictions for plant miRNAs in a controlled manner. We have implemented it in PAREfirst, a freely available software that can be used to predict functional miRNAs. As more sequenced genomes of different species become available, we hope that our tool will play an important role in the understanding of biology and evolution through the annotation of novel miRNAs and their functions.

ACKNOWLEDGMENT

The authors are grateful for Joshua Thody and Simon Moxon for their useful suggestions and contributions. The authors also appreciate the comments and suggestions from colleagues that improved the clarity of this paper. S. A. thanks the Saudi Cultural Bureau in London for funding her studentship. The authors declare that they have no conflicting interests.

REFERENCES

- [1] M. J. Aukerman and H. Sakai, "Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes," *Plant Cell*, vol. 15, no. 11, pp. 2730–2741, Nov. 2003, doi: 10.1105/tpc.016238.
- [2] X. Chen, "A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development," *Science*, vol. 303, no. 5666, pp. 2022–2025, Mar. 2004, doi: 10.1126/science.1088060.
- [3] M. W. Jones-Rhoades and D. P. Bartel, "Computational identification of plant microRNAs and their targets, including a stress-induced miRNA," *Mol. Cell*, vol. 14, no. 6, pp. 787–799, Jun. 2004, doi: 10.1016/j.molcel.2004.05.027.
- [4] A. L. Eamens, N. A. Smith, S. J. Curtin, M.-B. Wang, and P. M. Waterhouse, "The Arabidopsis thaliana double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes," *RNA*, vol. 15, no. 12, pp. 2219–2235, Dec. 2009, doi: 10.1261/rna.1646909.
- [5] Y. Tomari *et al.*, "RISC assembly defects in the Drosophila RNAi mutant armitage," *Cell*, vol. 116, no. 6, pp. 831–841, Mar. 2004.
- [6] C. Llave, K. D. Kasschau, M. A. Rector, and J. C. Carrington, "Endogenous and Silencing-Associated Small RNAs in Plants," *Plant Cell*, vol. 14, no. 7, pp. 1605–1619, Jul. 2002, doi: 10.1105/tpc.003210.
- [7] M. A. German *et al.*, "Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends," *Nat. Biotechnol.*, vol. 26, no. 8, pp. 941–946, Aug. 2008, doi: 10.1038/nbt1417.
- [8] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, Jan. 2004.
- [9] B. J. Reinhart, "MicroRNAs in plants," *Genes Dev.*, vol. 16, no. 13, pp. 1616–1626, Jul. 2002, doi: 10.1101/gad.1004402.
- [10] S. E. Schauer, S. E. Jacobsen, D. W. Meinke, and A. Ray, "DICER-LIKE1: blind men and elephants in Arabidopsis development," *Trends Plant Sci.*, vol. 7, no. 11, pp. 487–491, Nov. 2002.
- [11] Z. Xie, E. Allen, N. Fahlgren, A. Calamar, S. A. Givan, and J. C. Carrington, "Expression of Arabidopsis MIRNA Genes," *Plant Physiol.*, vol. 138, no. 4, pp. 2145–2154, Aug. 2005, doi:

- 10.1104/pp.105.062943.
- [12] J. T. Cuperus, N. Fahlgren, and J. C. Carrington, "Evolution and functional diversification of MIRNA genes," *Plant Cell*, vol. 23, no. 2, pp. 431–442, Feb. 2011, doi: 10.1105/tpc.110.082784.
- [13] B. C. Meyers *et al.*, "Criteria for Annotation of Plant MicroRNAs," *Plant Cell*, vol. 20, no. 12, pp. 3186–3190, Dec. 2008, doi: 10.1105/tpc.108.064311.
- [14] E. R. Mardis, "Next-generation DNA sequencing methods," *Annu. Rev. Genomics Hum. Genet.*, vol. 9, pp. 387–402, 2008, doi: 10.1146/annurev.genom.9.081307.164359.
- [15] W. J. Ansorge, "Next-generation DNA sequencing techniques," *New Biotechnol.*, vol. 25, no. 4, pp. 195–203, Apr. 2009, doi: 10.1016/j.nbt.2008.12.009.
- [16] E. Zhu *et al.*, "mirTools: microRNA profiling and discovery based on high-throughput sequencing," *Nucleic Acids Res.*, vol. 38, no. suppl_2, pp. W392–W397, Jul. 2010, doi: 10.1093/nar/gkq393.
- [17] W. Kang and M. R. Friedländer, "Computational Prediction of miRNA Genes from Small RNA Sequencing Data," *Front. Bioeng. Biotechnol.*, vol. 3, 2015, doi: 10.3389/fbioe.2015.00007.
- [18] M. J. Axtell and B. C. Meyers, "Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data," *Plant Cell*, vol. 30, no. 2, pp. 272–284, Feb. 2018, doi: 10.1105/tpc.17.00851.
- [19] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D140–D144, Jan. 2006, doi: 10.1093/nar/gkj112.
- [20] Z. Guo *et al.*, "PmiREN: a comprehensive encyclopedia of plant miRNAs," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1114–D1121, 08 2020, doi: 10.1093/nar/gkz894.
- [21] J. An, J. Lai, A. Sajjanhar, M. L. Lehman, and C. C. Nelson, "miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data," *BMC Bioinformatics*, vol. 15, no. 1, p. 275, Aug. 2014, doi: 10.1186/1471-2105-15-275.
- [22] C. Paicu *et al.*, "miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets," *Bioinformatics*, vol. 33, no. 16, pp. 2446–2454, Aug. 2017, doi: 10.1093/bioinformatics/btx210.
- [23] X. Yang and L. Li, "miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants," *Bioinformatics*, vol. 27, no. 18, pp. 2614–2615, Sep. 2011, doi: 10.1093/bioinformatics/btr430.
- [24] R. S. Taylor, J. E. Tarver, A. Foroozani, and P. C. J. Donoghue, "MicroRNA annotation of plant genomes – Do it right or not at all," *BioEssays*, vol. 39, no. 2, p. 1600113, 2017, doi: 10.1002/bies.201600113.
- [25] D. Swarbreck *et al.*, "The Arabidopsis Information Resource (TAIR): gene structure and function annotation," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D1009–1014, Jan. 2008, doi: 10.1093/nar/gkm965.
- [26] J. Thody, L. Folkes, Z. Medina-Calzada, P. Xu, T. Dalmay, and V. Moulton, "PAREsnip2: a tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules," *Nucleic Acids Res.*, vol. 46, no. 17, pp. 8730–8739, 28 2018, doi: 10.1093/nar/gky609.
- [27] T. Z. Berardini *et al.*, "The Arabidopsis information resource: Making and mining the 'gold standard' annotated reference plant genome," *Genes. N. Y. N* 2000, vol. 53, no. 8, pp. 474–485, Aug. 2015, doi: 10.1002/dvg.22877.
- [28] M. B. Stocks *et al.*, "The UEA sRNA Workbench (version 4.4): a comprehensive suite of tools for analyzing miRNAs and sRNAs," *Bioinforma. Oxf. Engl.*, vol. 34, no. 19, pp. 3382–3384, 01 2018, doi: 10.1093/bioinformatics/bty338.
- [29] N. Fahlgren and J. C. Carrington, "miRNA Target Prediction in Plants," in *Plant MicroRNAs: Methods and Protocols*, B. C. Meyers and P. J. Green, Eds. Totowa, NJ: Humana Press, 2010, pp. 51–57. doi: 10.1007/978-1-60327-005-2_4.
- [30] C. Brousse, Q. Liu, L. Beauclair, A. Deremetz, M. J. Axtell, and N. Bouché, "A non-canonical plant microRNA target site," *Nucleic Acids Res.*, vol. 42, no. 8, pp. 5270–5279, Apr. 2014, doi: 10.1093/nar/gku157.
- [31] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.
- [32] S. X. Ge, E. W. Son, and R. Yao, "iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data," *BMC Bioinformatics*, vol. 19, no. 1, p. 534, Dec. 2018, doi: 10.1186/s12859-018-2486-6.
- [33] N. Jiang *et al.*, "Synergy between the anthocyanin and RDR6/SGS3/DCL4 siRNA pathways expose hidden features of Arabidopsis carbon metabolism," *Nat. Commun.*, vol. 11, no. 1, Art. no. 1, May 2020, doi: 10.1038/s41467-020-16289-3.
- [34] K. Prüfer, U. Stenzel, M. Dannemann, R. E. Green, M. Lachmann, and J. Kelso, "PatMaN: rapid alignment of short sequences to large databases," *Bioinformatics*, vol. 24, no. 13, pp. 1530–1531, Jul. 2008, doi: 10.1093/bioinformatics/btn223.
- [35] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Malaysia; Pearson Education Limited, 2016. Accessed: Sep. 26, 2019. [Online]. Available: http://thuvinen-so.thanglong.edu.vn/handle/DHTL_123456789/4010
- [36] E. Allen, Z. Xie, A. M. Gustafson, and J. C. Carrington, "microRNA-Directed Phasing during Trans-Acting siRNA Biogenesis in Plants," *Cell*, vol. 121, no. 2, pp. 207–221, Apr. 2005, doi: 10.1016/j.cell.2005.04.004.
- [37] R. Lorenz *et al.*, "ViennaRNA Package 2.0," *Algorithms Mol. Biol.*, vol. 6, no. 1, p. 26, Nov. 2011, doi: 10.1186/1748-7188-6-26.
- [38] Z. Kuang, Y. Wang, L. Li, and X. Yang, "miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants," *Bioinformatics*, vol. 35, no. 14, pp. 2521–2522, Jul. 2019, doi: 10.1093/bioinformatics/bty972.
- [39] L. Folkes *et al.*, "PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing," *Nucleic Acids Res.*, vol. 40, no. 13, p. e103, Jul. 2012, doi: 10.1093/nar/gks277.
- [40] M. D. Saçar, H. Hamzeiy, and J. Allmer, "Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins?," *J. Integr. Bioinforma.*, vol. 10, no. 2, p. 215, Mar. 2013, doi: 10.2390/biecoll-jib-2013-215.
- [41] W. Zhang *et al.*, "Multiple distinct small RNAs originate from the same microRNA precursors," *Genome Biol.*, vol. 11, no. 8, p. R81, 2010, doi: 10.1186/gb-2010-11-8-r81.
- [42] C. Zeng, J. Xia, X. Chen, Y. Zhou, M. Peng, and W. Zhang, "MicroRNA-like RNAs from the same miRNA precursors play a role in cassava chilling responses," *Sci. Rep.*, vol. 7, no. 1, pp.

- 1–9, Dec. 2017, doi: 10.1038/s41598-017-16861-w.
- [43] B. Khraiweh *et al.*, “Transcriptional Control of Gene Expression by MicroRNAs,” *Cell*, vol. 140, no. 1, pp. 111–122, Jan. 2010, doi: 10.1016/j.cell.2009.12.023.
- [44] D. Baulcombe, “RNA silencing in plants,” *Nature*, vol. 431, no. 7006, pp. 356–363, Sep. 2004, doi: 10.1038/nature02874.
- [45] Tomato Genome Consortium, “The tomato genome sequence provides insights into fleshy fruit evolution,” *Nature*, vol. 485, no. 7400, pp. 635–641, May 2012, doi: 10.1038/nature11119.
- [46] S. Lopez-Gomollon, I. Mohorianu, G. Szittyta, V. Moulton, and T. Dalmay, “Diverse correlation patterns between microRNAs and their targets during tomato fruit development indicates different modes of microRNA actions,” *Planta*, vol. 236, no. 6, pp. 1875–1887, Dec. 2012, doi: 10.1007/s00425-012-1734-7.
- [47] N. Fernandez-Pozo *et al.*, “The Sol Genomics Network (SGN)-from genotype to phenotype to breeding,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D1036–1041, Jan. 2015, doi: 10.1093/nar/gku1195.
- [48] X. Song *et al.*, “Roles of DCL4 and DCL3b in rice phased small RNA biogenesis,” *Plant J. Cell Mol. Biol.*, vol. 69, no. 3, pp. 462–474, Feb. 2012, doi: 10.1111/j.1365-313X.2011.04805.x.
- [49] L. Wu, Q. Zhang, H. Zhou, F. Ni, X. Wu, and Y. Qi, “Rice MicroRNA effector complexes and targets,” *Plant Cell*, vol. 21, no. 11, pp. 3421–3435, Nov. 2009, doi: 10.1105/tpc.109.070938.
- [50] H. Sakai *et al.*, “Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics,” *Plant Cell Physiol.*, vol. 54, no. 2, p. e6, Feb. 2013, doi: 10.1093/pcp/pcs183.
- [51] Y. Kawahara *et al.*, “Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data,” *Rice N. Y. N.*, vol. 6, no. 1, p. 4, Feb. 2013, doi: 10.1186/1939-8433-6-4.
- [52] P. Kerpedjiev, S. Hammer, and I. L. Hofacker, “Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams,” *Bioinformatics*, vol. 31, no. 20, pp. 3377–3379, Oct. 2015, doi: 10.1093/bioinformatics/btv372.
- [53] P. May *et al.*, “The effects of carbon dioxide and temperature on microRNA expression in *Arabidopsis* development,” *Nat. Commun.*, vol. 4, p. 2145, Jul. 2013, doi: 10.1038/ncomms3145.
- [54] X. Wang *et al.*, “Degradation of unmethylated miRNA/miRNA*s by a DEDDy-type 3' to 5' exonuclease Atrimmer 2 in *Arabidopsis*,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 28, pp. E6659–E6667, Jul. 2018, doi: 10.1073/pnas.1721917115.
- [55] V. Pantaleo *et al.*, “Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis,” *Plant J. Cell Mol. Biol.*, vol. 62, no. 6, pp. 960–976, Jun. 2010, doi: 10.1111/j.0960-7412.2010.04208.x.
- [56] C. Llave, Z. Xie, K. D. Kasschau, and J. C. Carrington, “Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA,” *Science*, vol. 297, no. 5589, pp. 2053–2056, Sep. 2002, doi: 10.1126/science.1076311.
- [57] M. A. German *et al.*, “Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends,” *Nat. Biotechnol.*, vol. 26, no. 8, pp. 941–946, Aug. 2008, doi: 10.1038/nbt1417.
- [58] R. Wang *et al.*, “Construction and analysis of degradome-dependent microRNA regulatory networks in soybean,” *BMC Genomics*, vol. 20, no. 1, p. 534, Jun. 2019, doi: 10.1186/s12864-019-5879-7.
- [59] D. Yue, H. Liu, and Y. Huang, “Survey of Computational Algorithms for MicroRNA Target Prediction,” *Curr. Genomics*, vol. 10, no. 7, pp. 478–492, Nov. 2009, doi: 10.2174/138920209789208219.
- [60] C. Addo-Quaye, W. Miller, and M. J. Axtell, “CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets,” *Bioinformatics*, vol. 25, no. 1, pp. 130–131, Jan. 2009, doi: 10.1093/bioinformatics/btn604.
- [61] Y. Zheng, Y.-F. Li, R. Sunkar, and W. Zhang, “SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants,” *Nucleic Acids Res.*, vol. 40, no. 4, p. e28, Feb. 2012, doi: 10.1093/nar/gkr1092.
- [62] J. Zhang *et al.*, “Profiling of MicroRNAs and Their Targets in Roots and Shoots Reveals a Potential miRNA-Mediated Interaction Network in Response to Phosphate Deficiency in the Forestry Tree *Betula luminifera*,” *Front. Genet.*, vol. 12, p. 552454, 2021, doi: 10.3389/fgene.2021.552454.
- [63] X. Ma *et al.*, “Small RNA and Degradome Deep Sequencing Reveals the Roles of microRNAs in Seed Expansion in Peanut (*Arachis hypogaea* L.),” *Front. Plant Sci.*, vol. 9, p. 349, 2018, doi: 10.3389/fpls.2018.00349.
- [64] S. Ravichandran, R. Ragupathy, T. Edwards, M. Domaratzki, and S. Cloutier, “MicroRNA-guided regulation of heat stress response in wheat,” *BMC Genomics*, vol. 20, no. 1, p. 488, Jun. 2019, doi: 10.1186/s12864-019-5799-6.
- [65] A. T. Djami-Tchatchou, N. Sanan-Mishra, K. Ntushelo, and I. A. Dubery, “Functional Roles of microRNAs in Agronomically Important Plants-Potential as Targets for Crop Improvement and Protection,” *Front. Plant Sci.*, vol. 8, p. 378, 2017, doi: 10.3389/fpls.2017.00378.
- [66] Y.-F. Li *et al.*, “An improved method of constructing degradome library suitable for sequencing using Illumina platform,” *Plant Methods*, vol. 15, p. 134, 2019, doi: 10.1186/s13007-019-0524-7.
- [67] S.-S. Lin, Y. Chen, and M.-Y. J. Lu, “Degradome Sequencing in Plants,” *Methods Mol. Biol. Clifton NJ*, vol. 1932, pp. 197–213, 2019, doi: 10.1007/978-1-4939-9042-9_15.
- [68] M.-C. Carpentier, C. Bousquet-Antonelli, and R. Merret, “Fast and Efficient 5'P Degradome Library Preparation for Analysis of Co-Translational Decay in *Arabidopsis*,” *Plants Basel Switz.*, vol. 10, no. 3, Mar. 2021, doi: 10.3390/plants10030466.
- [69] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, “miRBase: from microRNA sequences to function,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D155–D162, Jan. 2019, doi: 10.1093/nar/gky1141.



Salma Alzahrani received the bachelor's degree in computing and information technology from King Abdulaziz University, Saudi Arabia, in 2014, and the MSc degree in computing sciences from University of East Anglia, United Kingdom, in 2015. Currently, she is working towards the PhD degree in computing sciences in University of East Anglia. Her research interests include bioinformatics and software engineering.



Tamas Dalmay graduated from the University of Budapest and obtained the Ph.D. degree in molecular virology in Hungary. He moved to The Sainsbury Laboratory (Norwich, UK) with an EMBO fellowship in 1995 to work on the genetics of gene silencing. He has been group leader at the University of East Anglia, School of Biological Sciences (Norwich, UK) since 2002 where his research group has been working on microRNAs (miRNAs) and other non-coding short RNAs. He has been the Head of School since 2014.



Christopher Applegate received a BSc degree in Applied Computing Science from the University of East Anglia, UK in 2009, and a PhD from the University of East Anglia, UK in 2013. In his doctoral thesis, he investigated the challenges related to the design and simulation of traffic networks for use in virtual environments. Between 2013 and 2016, he worked as a postdoc at the University of East Anglia, UK on projects which involved developing a computational tool to semi-automatically generate tailored immobilisation shells for patients undergoing radiotherapy treatment from CT scans, and bioinformatics software to help predict functional microRNA. From 2016-2017, he worked as a Scientific Programmer at Earlham Institute, UK where he developed novel high-throughput image processing algorithms for applications in agri-tech and crop research.



Leighton Folkes completed the undergraduate studies in Computing Science at the University of East Anglia (UEA, Norwich, UK) and continued at the UEA to complete the Ph.D. degree in Bioinformatics, graduating in 2014. He then worked as a Post-Doctoral Scientist at the Earlham Institute. In 2016 he rejoined the UEA where he is employed as a Senior Research Associate (Bioinformatician). His research interests include the design and development of computational methods for the analysis of next generation sequencing data.

aply treatment from CT scans, and bioinformatics software to help predict functional microRNA. From 2016-2017, he worked as a Scientific Programmer at Earlham Institute, UK where he developed novel high-throughput image processing algorithms for applications in agri-tech and crop research.



David Swarbreck received a BSc degree in Genetics (Hons) from the University of Liverpool, UK in 1998 and completed a Ph. D degree at Rothamsted Research and the University of Bristol in 2002. He then worked as a Senior computer biologist at the Wellcome Trust Sanger Institute, Hinxton before joining The Arabidopsis Information Resource (TAIR) Carnegie Institute, USA in 2005. He moved to the Earlham Institute in 2010 where he is currently Head of Core Bioinformatics, his group are actively engaged in a diverse range of collaborative projects including international efforts in plant and insect genomics and develop software and analysis pipelines that support the Institute's research activities.



Vincent Moutlon completed the undergraduate studies at University of Warwick in 1987. After that, he moved to the USA, where he completed an M.Sc degree at University of Washington in 1991, and the Ph.D. degree at Duke University in 1994. He then worked as a researcher at University of Bielefeld, DE, and University of Canterbury/Massey University, NZ. In 1997 he moved to Sweden, where he was employed as Senior Lecturer at Mid Sweden University and, as of 2002, Professor in Bioinformatics at Uppsala University. He joined University of East Anglia in 2004 where he is employed as Professor in Computational Biology. His research interests include phylogenetics, computational biology of RNA, short RNAs, metatranscriptomics, algorithms in bioinformatics, and discrete structures such as graphs and finite metric spaces.

tively engaged in a diverse range of collaborative projects including international efforts in plant and insect genomics and develop software and analysis pipelines that support the Institute's research activities.