Prediction and Characterization of miRNA/Target Pairs in Non-Model Plants Using RNA-seq

Kira C. M. Neller,[1] Alexander Klenov,[1] and Katalin A. Hudak[1,2]

[1]Department of Biology, York University, Toronto, Ontario, Canada
[2]Corresponding author: *hudak@yorku.ca*

Plant microRNAs (miRNAs) are ∼20- to 24-nucleotide small RNAs that post-transcriptionally regulate gene expression of mRNA targets. Here, we present a workflow to characterize the miRNA transcriptome of a non-model plant, focusing on miRNAs and targets that are differentially expressed under one experimental treatment. We cover RNA-seq experimental design to create paired small RNA and mRNA libraries and perform quality control of raw data, de novo mRNA transcriptome assembly and annotation, miRNA prediction, differential expression, target identification, and functional enrichment analysis. Additionally, we include validation of differential expression and miRNA-induced target cleavage using qRT-PCR and modified RNA ligase–mediated 5′ rapid amplification of cDNA ends, respectively. Our procedure relies on freely available software and web resources. It is intended for users that lack programming skills but can navigate a command-line interface. To enable an understanding of formatting requirements and anticipated results, we provide sample RNA-seq data and key input/output files for each stage. © 2019 The Authors. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Keywords: de novo transcriptome assembly • differential expression • miRNA prediction • miRNA target • non-model plant • RNA-seq • small RNA

INTRODUCTION

Plant small RNAs are ∼20 to 24 nucleotides (nt) in length and are classified as either microRNAs (miRNAs) or small interfering RNAs (siRNAs), depending on their biogenesis. miRNAs are derived from single-stranded mRNA precursors (pre-miRNAs) that adopt a characteristic hairpin structure. The biological relevance of an miRNA is defined by the functional role of its mRNA target. Plant miRNAs tend to have high sequence complementarity with targets and act by inducing transcript cleavage, resulting in mRNA decay. This differs from animal miRNAs, which tend to share a smaller 'seed' region of complementarity with targets and act through translational inhibition. By modulating

CURRENT PROTOCOLS
A Wiley Brand
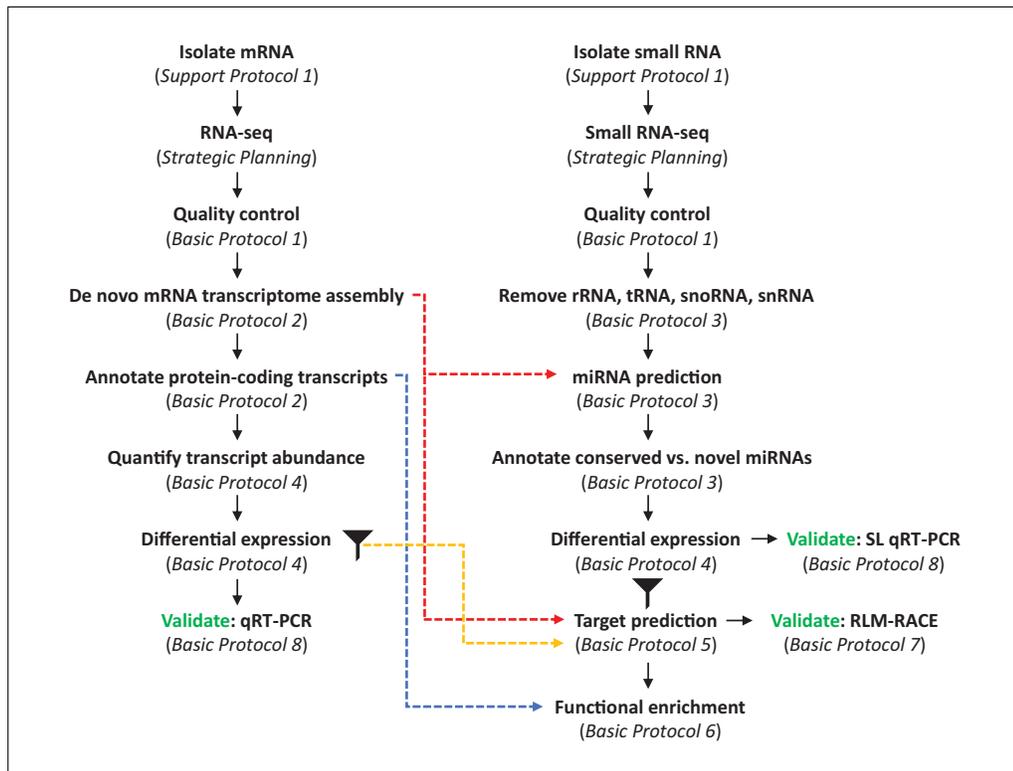
Figure 1 Overview of the article. Colored arrows indicate points of integration between miRNA and target analysis. Differential expression analysis is applied as a filtration step to identify the most biologically relevant miRNA/target pairs.

the level of transcript abundance or protein accumulation, miRNAs provide an important layer of gene expression regulation.

This article outlines the steps involved in large-scale miRNA prediction and characterization, with a focus on non-model plants that lack genomic references. The workflow of the protocols in this article is provided in Figure 1. We cover strategic planning, including RNA-seq experimental design and considerations for library preparation and sequencing; quality control of raw small RNA and mRNA reads (Basic Protocol 1); de novo mRNA transcriptome assembly and annotation of protein-coding genes through homology and protein domain identification (Basic Protocol 2); miRNA prediction and identification of conserved versus novel sequences (Basic Protocol 3); miRNA and mRNA differential expression analysis (Basic Protocol 4); prediction of miRNA/target pairs (Basic Protocol 5); and gene ontology (GO) term functional enrichment analysis (Basic Protocol 6). We also include wet-lab methods to validate bioinformatic predictions derived from the above analyses: qRT-PCR and stem-loop qRT-PCR to confirm mRNA and miRNA differential expression, respectively (Basic Protocol 8), and modified RNA ligase–mediated 5′ rapid amplification of cDNA ends (RLM-RACE) to detect miRNA-induced cleavage of target transcripts (Basic Protocol 7). This protocol enables the detection of miRNAs that act through transcript cleavage rather than translational inhibition of targets. Although both mechanisms of miRNA activity exist in plants, the former is more prevalent.

This protocol is based on our previous work to identify defense-associated miRNA/target pairs in the non-model plant *Phytolacca americana* (Neller, Klenov, & Hudak, 2016, 2018). To provide greater context for the reader, we carry out the bioinformatic workflow using an example RNA-seq dataset that was generated by random sampling of our published data. We aim to be user-friendly, introducing graphical interface options when available. However, we assume that the reader can navigate a UNIX-like operating system

from a command-line interface, since this is commonly required to conduct large-scale computational analysis.

STRATEGIC PLANNING

Experimental design

The method described in this article is based on small RNA and mRNA samples that are both derived from one control and one experimental condition. The sample data that we provide were used to characterize miRNAs associated with defense in *P. americana*. In this case, plants were sprayed with either ethanol (control) or jasmonic acid (JA, experimental treatment). JA is a plant hormone that mediates plant defense against insect herbivores and necrotrophic pathogens. It is possible to follow this procedure without incorporating a conditional variable; however, we advocate the use of differential expression as a filter to identify the most biologically meaningful miRNA/target pairs.

To reduce the impact of experimental variability on miRNA and target expression, we suggest the use of paired small RNA and mRNA samples. This means that for each sample, both the small RNA and mRNA fractions are extracted. Additionally, the use of biological replicates (i.e., different plants that undergo the same treatment) is recommended for robust differential expression analysis. Due to the expense of RNA-seq, technical replicates (i.e., identical samples) are not usually included in the experimental design. Furthermore, it is common to pool equal volumes of RNA from independent plants undergoing the same treatment. Such 'pooled biological replicates' deviate from the definition of true biological replicates, but ease downstream analysis. Specifically, this strategy increases the number of plants tested (leading to greater biological representation) while reducing technical variability across replicates. The downside is that it does not detect differences among independent plants. However, both biological and technical variability may be assessed through the more economical method of qRT-PCR.

Our study comprised two conditions (control versus JA), with three pooled biological replicates per condition (i.e., six total samples). Each pooled replicate consisted of equal amounts of RNA from three independent plants; therefore, a total of 18 plants were used. Both small RNA samples and mRNA-enriched samples were prepared, arranged as follows, with replicate number indicated: $(C1_{sRNA}: C1_{mRNA})$, $(C2_{sRNA}: C2_{mRNA})$, $(C3_{sRNA}: C3_{mRNA})$, $(J1_{sRNA}: J1_{mRNA})$, $(J2_{sRNA}: J2_{mRNA})$, $(J3_{sRNA}: J3_{mRNA})$. Here, 'C' and 'J' indicate control and JA conditions, respectively, and 'sRNA' and 'mRNA' indicate small RNA and mRNA fractions, respectively. Our sampling design is illustrated in Figure 2.

RNA extraction and enrichment of mRNA or small RNA

The sequencing facility will likely have different options for sample submission that vary in level of downstream processing and associated cost. Commonly accepted submission types include total RNA, enriched RNA (mRNA or small RNA), and ready-to-sequence libraries. For readers who wish to prepare their own small RNA libraries, we recommend the protocol by Mathioni, Kakrana, and Meyers (2017), which covers total RNA extraction from plant tissue, small RNA enrichment, and library construction. Given that very little RNA is required for sequencing (1 to 2 µg of total RNA and 500 ng to 1 µg of small RNA) and that multiple samples need to be processed, the use of RNA isolation kits may be preferred for RNA extraction. We recommend contacting the facility to discuss specific requirements for sample preparation and any recommended kits.

Prior to RNA-seq library preparation, either mRNA enrichment or rRNA depletion of total RNA must be performed. Without this step, most reads would derive from rRNA due to its high relative abundance. The decision to use one strategy over the other depends
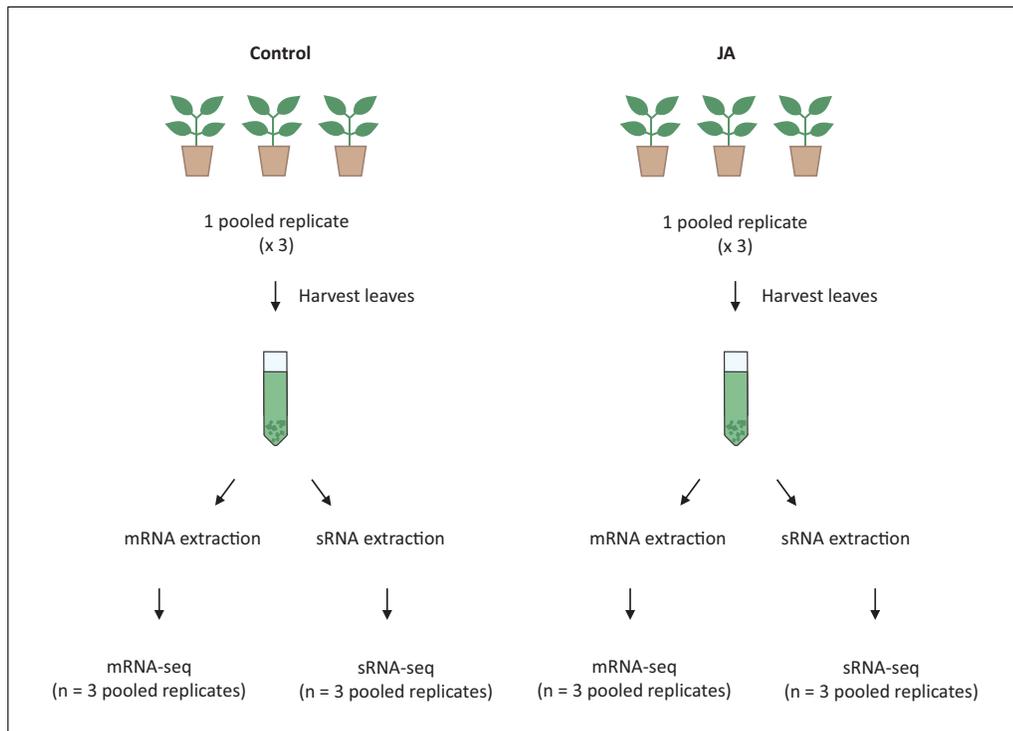
Figure 2 Generation of paired mRNA and small RNA (sRNA) samples for RNA-seq. Sample data provided in this article are from Neller et al. (2018). Our study included two conditions: an ethanol control treatment and jasmonic acid (JA) experimental treatment. For each condition, three pooled biological replicates were prepared, with each replicate consisting of equal amounts of RNA from three independent plants. Both the sRNA and mRNA fractions were extracted for each plant to produce a paired sRNA:mRNA design. In total, 18 plants were used (3 plants per pooled replicate × 3 pooled replicates × 2 conditions).

on the project goal. mRNA enrichment is based on the selection of RNA species containing a poly(A) tail and 5′ cap, whereas rRNA depletion selectively removes rRNA and consequently enriches for all other RNAs, including non-coding RNAs. rRNA depletion involves hybridization of probes containing conserved rRNA sequences, the probes being linked to magnetic beads to facilitate rRNA removal following hybridization. In addition to cytoplasmic rRNA, kits often include probes to remove chloroplast and mitochondrial rRNA. The benefit of rRNA depletion over mRNA selection is that the former results in more information retention. In terms of small RNA characterization, rRNA depletion would enable the identification of both siRNA and miRNA precursors, since the former tend to lack poly(A) tails. We chose mRNA enrichment for our study; not only was it more economical, it was the more conservative approach given the lack of known rRNA sequences for our non-model plant.

Small RNA enrichment involves isolation of the low-weight RNA fraction. As detailed by Mathioni et al. (2017), this can be accomplished by separating total RNA on a urea-polyacrylamide gel and excising the small RNA fraction. This was the chosen strategy for our experiment. As an alternative option, the small RNA fraction can be obtained using dedicated spin kits that employ size-exclusion filters. Such kits are especially convenient in facilitating a paired study design, as one can quickly obtain total RNA, small RNA–enriched, and small RNA–depleted fractions from a single sample.

Library preparation and sequencing

The final step prior to RNA-seq is the preparation of sequence libraries, one per sample. With the mRNA-enriched or rRNA-depleted fraction as input, standard library preparation includes the following steps: RNA fragmentation, cDNA synthesis, adapter ligation,

and PCR amplification. For RNA-seq, we recommend the preparation of paired-end, strand-specific libraries. In a paired-end library, the ligated adapters contain an additional priming site to enable sequencing of both ends of the DNA fragment. Size selection is employed to ensure that the majority of DNA fragments are long enough to avoid sequence overlap between the two read-pairs. Although more expensive, paired-end reads provide greater information for downstream transcriptome assembly, since it is known that both pairs derive from a single fragment and are a certain distance apart. This helps resolve regions that are repeat-rich or of low complexity. We also recommend the use of strand-specific library preparation methods to retain information on the orientation of the originally sequenced transcript. Distinguishing between sense and anti-sense transcription of overlapping genes enables more accurate detection of differential expression. For our study, we used strand-specific sequencing for both mRNA and small RNA-seq. Paired-end sequencing was used only for mRNA-seq; it is not necessary for small RNA-seq because miRNAs are short enough to be sequenced completely with one pass.

It is also important to consider the balance between sequencing depth and number of experimental conditions/replicates. Since a single lane provides a fixed output of reads, increasing the number of samples per lane will reduce the amount of reads per sample. The strategy of combining multiple samples in a lane is termed 'multiplexing', whereby adapters with sample-specific indexes are used to distinguish independent libraries. Since RNA-seq has fixed costs both per lane and per sample, multiplexing can reduce the overall cost of sequencing at the expense of reduced coverage per sample. If this is the first RNA-seq study for your non-model plant, we suggest that depth take priority; the resulting high-coverage references will serve as a foundation for future work, enabling less deep sequencing in future experiments. In our study, two sequencing lanes were used for both small RNA-seq and mRNA-seq, with respective samples equally distributed between the lanes. The project goals should be discussed in advance with the sequencing facility, as they can provide an expected number of reads given the sequencing platform and number of lanes used.

We note here that two different sequencing platforms were used in our experiment. We chose SOLiD (commercially discontinued since 2016) for small RNA-seq because it had a lower error rate than Illumina at the time of the study, with the latter platform used for mRNA-seq. As miRNAs are ~20 to 24 nt in length, a low error rate ensures that more reads are retained during alignment steps. Given a fixed proportion of sequence errors, a shorter sequence is more difficult to align to a reference than a longer one. Presently, we suggest the use of Illumina sequencing to perform both mRNA-seq and small RNA-seq. However, we maintain the use of SOLiD data in this procedure because such data are widely available in public sequencing repositories, necessitating an understanding of their special considerations.

QUALITY CONTROL PROCESSING OF RNA-seq DATA

In this protocol, we cover the format of raw SOLiD and Illumina RNA-seq data, necessary file conversions, and steps to 'clean' the reads. Cleaning entails the removal of adapter sequences and low-quality bases. This protocol is designed to be carried out using a public Galaxy account. Galaxy is an open-source, web-based platform that enables researchers without command-line programming experience to perform basic bioinformatic analyses (Giardine et al., 2005). The utility of Galaxy lies with its graphical interface, which allows users to interact with their data, perform large-scale operations, and run standard programs, all without exposure to the underlying computational infrastructure. We introduce Galaxy at this stage to provide the reader with a visual sense of the input data prior to downstream analysis at the terminal. Sample data and final output files are provided in the accompanying Supporting Information.
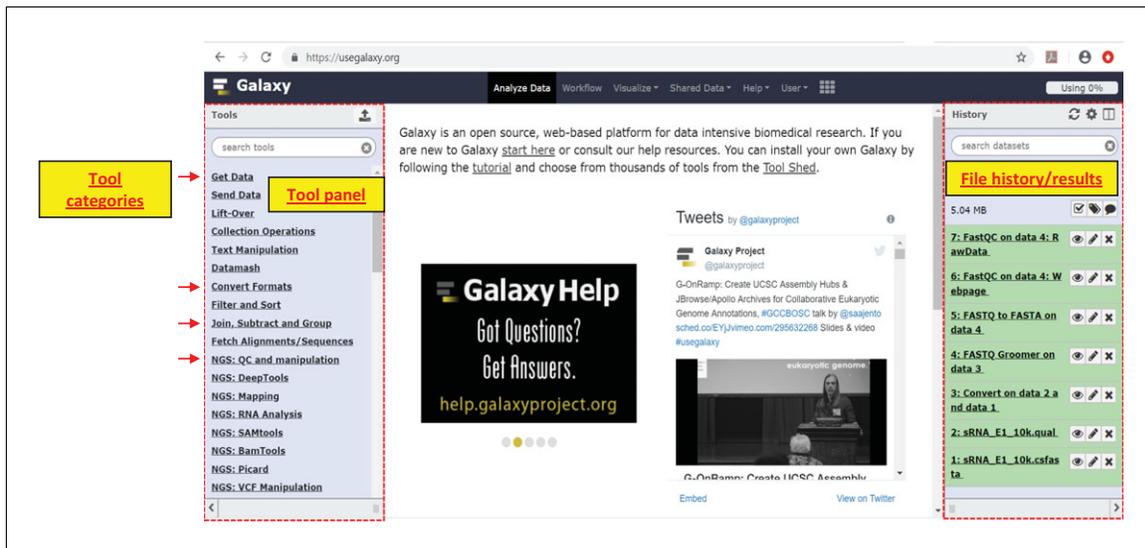
Figure 3 Layout of the Galaxy workspace. The tool panel is shown, with arrows indicating the location of tools used in this article. The history panel is also shown, containing user-uploaded files as well as output files.

Materials

A personal or desktop computer with internet access
A public Galaxy user account; register at: *https://usegalaxy.org/*
Raw SOLiD small RNA-seq dataset: `1.sRNA_reads/1.raw`
Raw Illumina mRNA-seq dataset: `2.mRNA_reads/1.raw`

Process the SOLiD small RNA-seq data

1. Upload the small RNA-seq files to Galaxy.

   To do this, first navigate to Galaxy at *https://usegalaxy.org/* and sign in with your account credentials. From the tool panel on the left, select the category 'Get Data'. Click on the tool 'Upload file from computer'. Select the option 'choose a local file' to browse your file directory. Upload the 12 small RNA data files. See Figure 3 for an annotated layout of the Galaxy workspace.

   Each small RNA sample consists of two individual files, with the extensions .qual and .csfasta. There are 12 total files (2 conditions × 3 replicates × 2 sequence files). The .qual file contains a quality score (also known as a Phred or Q score) for each sequenced base. A Q score is an integer ranging from 0 to 42 that is inversely associated with the error probability of the sequenced base, meaning that a base with a higher Q score is more likely to be correct than one with a lower score. The .csfasta file contains the sequenced reads in color space fasta format, where each di-nucleotide (i.e., two consecutive nucleotides) is represented as one of four possible colors (blue=0, green=1, yellow=2, and red=3). This so-called 2-base encoding derives from the ligation-based SOLiD sequencing technology, which differs from the sequencing-by-synthesis method used by Illumina.

   As an example, here is a single record (i.e., read) from one .csfasta file:

   ```
   >135_886_1067_F3
   T3232011323121012200013133020103031 3
   ```

   The name/identifier of the read is provided in the first line, following '>'; this one is designated 135_886_1067_F3 (with F3 indicating that these sequences are to be read in the forward direction, i.e., 5′ → 3′). The second line is the sequence in color space format, apart from the 'T', which is the last sequenced base of the primer.

Here is the same read from the corresponding `.qual` *file:*

```
>135_886_1067_F3
31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31
31 31 28 31 31 31 31 31 31 31 31 31 31 31 31 31 31
```

For each of the 35 sequenced bases in the `.csfasta` *file (not including the 'T'), there is an associated quality score in the* `.qual` *file.*

2. For each sample, combine the `.qual` and `.csfasta` files into a single fastq file.

    Under the tool category 'NGS: QC and manipulation', select 'Convert SOLiD output to fastq'. For the 'reads' option, provide the `.csfasta` file (e.g., `sRNA_E1_10k.csfasta`), and for the 'qualities' option, provide the corresponding `.qual` file for that sample (e.g., `sRNA_E1_10k.qual`). Set 'trim trailing F3 and R3?' to 'yes'.

    Now, the read above is represented in fastqcssanger format as follows:

```
@135_886_1067
T32320113231210122000131330201030313
+
@@@@@@@@@@@@@@@@@@@@=@@@@@@@@@@@@@
```

    In fastq format, each record consists of four lines. The identity of the read is defined in the header, following the @ sign. The sequence (still in color space here) is in the second line. The third line is a plus sign, which separates the sequence and quality information. The fourth line contains Q scores as before, but now they have been converted to corresponding ASCII (base 33) characters.

3. Decode the sequence from color space to base space.

    Use the tool 'FASTQ Groomer' under the category 'NGS: QC and manipulation'. For 'file to groom', select the fastqcssanger file generated from step 2. For 'input FASTQ quality scores type', select 'color space sanger'.

    Now the read is represented in standard fastq sanger format:

```
@135_886_1067
AGCTTGTAGCAGTTGAGGGGTACGCCTTGGCCGTA
+
@@@@@@@@@@@@@@@@@@@@=@@@@@@@@@@@@@
```

    Note that conversion from color space to base space was possible due to the known identity of the first base, i.e., the 'T' remaining from the primer sequence. With 2-base encoding, each color represents four possible di-nucleotides, so knowledge of the first base is essential for correct decoding. If you are working with SOLiD data that does not contain a known first base, then you must instead convert the reference sequence (i.e., genome or transcriptome) to color space for downstream alignments. However, since the SOLiD platform has been discontinued commercially since 2016, this option is not incorporated in all mapping tools, so be sure to select a mapper that includes this feature if required.

4. Obtain an overview of library quality.

    Run the 'FastQC' tool under the category 'NGS: QC and manipulation'. The file to be evaluated is the fastq output from step 3.

    FastQC evaluates multiple areas, including per-base quality scores, overrepresented sequences, duplicated sequences, adapter contamination, and kmer/GC content. Each

test result is associated with a green check mark (normal), orange triangle (slightly abnormal), or red cross (very unusual). The assumption is that normal libraries are random and diverse; however, some experiments or library preparation methods create expected biases, so an abnormal or failed result is not always cause for concern. For example, RNA-seq libraries that have incorporated random priming will tend to fail the 'per base sequence content' check, since random priming has inherent sequence bias. However, this does not impede downstream analysis.

5. Remove contaminating adapter sequences.

   Select the 'Clip' tool from the category 'NGS: QC and manipulation'. For input, select the fastq file output from step 3. Set 'minimum sequence length' to 18, 'custom clipping sequence' to CGCCTTGGCCGTACAGCAG, and both 'discard sequences with unknown bases' and 'output only clipped sequences' to 'yes'.

   Due to adapter read-through, which occurs when a sequenced fragment is shorter than the defined read length of sequencing, a portion of raw reads will include contamination of the form 5′-REALSEQUENCE<u>ADAPTER</u>-3′. This is solved by truncating reads at the first base of identified adapter sequence. If your RNA-seq data were obtained from a sequencing company, the company will provide you with the adapter associated with each library. If you are working with publicly available data, you may not have access to the adapter sequence. In this case, FastQC should have identified the adapter in the previous step, either as an 'overrepresented sequence' or a standard adapter.

   The remaining parameters are set to filter out reads unlikely to be derived from bona fide miRNAs. Since most miRNAs are 21 nt, we set the minimum length after adapter clipping to 18 nt. Similarly, given our read length of 35 bp, any read derived from a true miRNA should contain adapter read-through; therefore, we output only sequences that were clipped. Non-clipped reads are likely derived from longer products of RNA degradation. Finally, discarding bases with unknown nucleotides is important for downstream miRNA and target prediction, both of which rely on sequence alignment.

6. Remove low-quality bases from reads.

   Select 'Trimmomatic' from the category 'NGS: QC and manipulation'. Specify that the data are single-end reads and choose the adapter-trimmed fastq file produced from step 5 as input. Select 'sliding window trimming' as the operation to perform, with default parameters.

   As you may have observed from the per-base sequence quality plot in the FastQC report, Q values decline steadily toward the 3′ end of reads; this is due to a reduced signal-to-noise ratio with increased amplification. Some programs resolve this issue by truncating the read at the first base below a specified quality threshold. Trimmomatic offers a more flexible 'sliding window' approach, truncating the read once the average Q value (over four bases) is reduced to 20.

7. Apply a final length filter.

   Use the tool 'Filter FASTQ' from the category 'NGS: QC and manipulation'. Set 'maximum length' to 28 nt.

   This filter removes reads unlikely to be miRNAs.

8. Convert the fastq file to fasta format.

   Choose the tool 'FASTQ to FASTA' from the category 'Convert Formats'. For input, select the output fastq file from step 7. Set 'rename sequence names in output file' to 'no'.

   This is the processed file that will be used for downstream miRNA prediction. The conversion from fastq to fasta (i.e., removal of quality information) was made to conform with input requirements of downstream software.

The read that we were tracking above now looks like this:

```
>135_886_1067
AGCTTGTAGCAGTTGAGGGGTA
```

A comparison with the raw sequence reveals that the adapter was trimmed correctly.

Process the Illumina mRNA-seq data

9. Upload the mRNA-seq data to Galaxy.

   See the instructions provided in step 1.

   Each mRNA-seq sample has two paired fastq files, designated R1 or R2, indicating that they contain sequences in the forward or reverse-complement orientations, respectively. There are twelve mRNA-seq files (2 conditions × 3 replicates × 2 paired sequence files).

   A read from an R1 file looks like this:

```
@HWI-
D00423:99:HB533ADXX:2:1213:12544:78916 1:N:0:CGATGT
TTCCTGTAGGCTGGGAGATTGGAAAGGTACCATGCGATTAATGGGCTGGATGT
CACTGGGGTCTCAAGTCCTCCAATGAATGGGTCTCCATTCAAGGGCTGGATTA
CTTGGTATGTTTGCTTATCAGATTGAATGGCCTTGATGGTAAAGG
+
@CBFFFFHGHHHJJHIIJIJJIJJJJEFHIIJJJJIJJJJIJIJJJJJJJJJJ
IJIIJJJJHHHHHHHFFFFFFEEEEEEEDDDCDDDDDDFFEEDDDDDDDDDDD
DDDDDACDDEEEDDDDDDDEEDCDDDDDDDDDDDDDCDDDDEDDDD
```

   This format is identical to that above for the small RNA reads in fastq format. The only difference is that the header contains additional information.

   The paired read from the corresponding R2 file looks like this:

```
@HWI-
D00423:99:HB533ADXX:2:1213:12544:78916 2:N:0:CGATGT
CAACAACATCACCAATGGCCAGCCAATTAAAGAGTGGCTTCACCGCAACAATT
GTGACTCCCAAGGGCATCTCTGGTCCTGCCTTAAGGCCCTTCCCTTCCATCAG
GAGGCACCATTCCTTTACCATCAAGGCCATTCAATCTGATAAGCA
+
CCCFFFFFHHHHHJJJJJJJJJJJIJJJJIJJJGHJJJJJJJJIJJJJIJJJ
JIIJIIJJJHHGFFFFEEDEEEECCDDDDDDDDDDDDDDDDDDDDDDDDEDDD
DDDDDDDDDDEEDDDDDDDDDDDDDDDDDEEEEDDEDEDDDDC
```

10. Run *FastQC* on each fastq file to obtain an overview of quality.

    The 'FastQC' tool can be run as in step 4. Since the raw mRNA-seq reads are already in fastq format, no additional conversions are necessary.

11. Remove adapter sequences and low-quality bases, and apply a minimum length filter.

    Use 'Trimmomatic' from step 6. Specify that the reads are paired-end and provide the R1 and R2 files. Choose 'yes' to 'perform initial ILLUMINACLIP step'. Choose 'Truseq3 paired-end adapters' from the available adapter sequences. Add

two additional options: 'sliding window trimming' as before and 'drop reads below a specified read length' (default is 20 nt).

Since paired-end data are available, palindromic mode can be engaged in Trimmomatic for improved detection of partial adapter sequences, down to as little as one nucleotide. Another useful feature of Trimmomatic is that it maintains correspondence of paired-end reads; that is, if one member of the pair fails quality control, the other member of the pair is automatically discarded. Correspondence between paired-end files is often required for downstream software.

12. Download and save the clean small RNA-seq and mRNA-seq files.

At this point, you should have 18 total files (6 small RNA-seq fasta files and 12 paired mRNA-seq fastq files). To check that all steps were performed correctly, compare your results with the clean files `1.sRNA_reads/2.clean` *and* `2.mRNA_reads/2.clean` *(see Supporting Information).*

TRANSCRIPTOME ASSEMBLY AND ANNOTATION

This protocol covers de novo transcript assembly and annotation using Trinity and Trinotate, respectively. We combine the clean mRNA-seq reads of all samples to produce a single reference transcriptome. From the assembled transcripts, the most likely open reading frames (ORFs) and their translated protein sequences are obtained using TransDecoder. The transcript and protein sequences are used in BLASTX and BLASTP homology searches, respectively, against the curated SwissProt protein database. Additionally, the protein sequences are searched for conserved Pfam protein domains. Using Trinotate, the search results are built into a comprehensive annotation report that provides insight into the function of each transcript. All input and output files available in the Supporting Information are indicated in the code boxes with green and red font, respectively.

Materials

A computer or high-performance cluster running a UNIX-based operating system
Clean mRNA-seq dataset from Basic Protocol 1: `2.mRNA_reads/2.clean`
Sample information file: `3.trinity/samples.txt`
The following programs will be run in the terminal through the command line. They must be installed according to the guidelines at the provided URLs, with the base installation directory accessible from the user's `$PATH`:
Trinity (v. 2.8.4): *https://github.com/trinityrnaseq/trinityrnaseq/wiki*
TransDecoder (v. 5.5.0): *https://github.com/TransDecoder/TransDecoder/wiki*
Trinotate (v. 3.1.1): *https://github.com/Trinotate/Trinotate.github.io/wiki*
NCBI BLAST+ (v. 2.7.1): *http://www.ncbi.nlm.nih.gov/books/NBK52640/*
HMMER (v. 3.2.1): *http://hmmer.org/*

Assemble the mRNA transcriptome

1. Use Trinity to de novo assemble the clean mRNA-seq reads.

To do this, type the following command into the terminal:

```
Trinity --seqType fq --max_memory 50G --CPU 16
--samples_file samples.txt --SS_lib_type RF
```

Here, we have specified the format of input sequence data (fastq), the maximum allowable memory (50 Gb), the number of available threads for computing on our system (16), the availability of a file containing sample information (`samples.txt` *), and the orientation of the strand-specific, paired-end reads (RF). The latter parameter means that for each*

pair, the R1 read will be reverse-complemented, while the R2 read will be processed as-is in the forward direction. Note that the relationship between read-pairs depends on the strategy employed during strand-specific library preparation.

The 'samples' file is a tab-delimited text file specifying the location of clean mRNA-seq reads and information about biological replicates for downstream analysis. Note that the paths must be modified to correspond with your own file system. The `samples.txt` *file looks like this:*

```
E E1 /path/to/mRNA_E1_10k_R1.fastq /path/to/mRNA_E1_
10k_R2.fastq
E E2 /path/to/mRNA_E2_10k_R1.fastq /path/to/mRNA_E2_
10k_R2.fastq
E E3 /path/to/mRNA_E3_10k_R1.fastq /path/to/mRNA_E3_
10k_R2.fastq
J J1 /path/to/mRNA_J1_10k_R1.fastq /path/to/mRNA_J1_
10k_R2.fastq
J J2 /path/to/mRNA_J2_10k_R1.fastq /path/to/mRNA_J2_
10k_R2.fastq
J J3 /path/to/mRNA_J3_10k_R1.fastq /path/to/mRNA_J3_
10k_R2.fastq
```

The first column denotes the treatment group (E or J) that the sample belongs to; the second column provides the sample replicate number (1, 2, or 3); and the third and fourth columns provide file locations of the R1 and R2 reads, respectively, for each sample.

Upon completion, the assembled transcriptome will be available as `Trinity.fasta` *in your Trinity output directory.*

Inspecting `Trinity.fasta` *reveals ~3000 assembled transcripts. An example sequence looks like this:*

```
>TRINITY_DN66_c0_g1_i1 len=299 path=[0:0-99 2:100-
298]
CATTACTCCCAACGTCCATCTCTCTCCGAAGCTCTTTCTCTCTCTTCAACTCC
AAGTTCTTTACCTTTTAAGAGAGAGGAAAATGGAGGCCATGAGGATTAGGGTT
GCATTCGCCGTCGTGATGATGCTCATCGCCGTCTCCACCGTCCAGAACGCCGC
CGCCGCCGATGCTCCGGCGCCGGCACCGACCTCCGACGCCGCCACCTTCCTCC
CCGCCGCATTTGTCTCCCTCGTCGCCCTAGCATTCGGCTTTCTCTTCTGAATT
TCGAGATTCTCCGTTACCTTTTGTTCAATTTTTG
```

The unique identifier of the sequence is '`TRINITY_DN66_c0_g1_i1`*', where '*`c`*', '*`g`*', and '*`i`*' indicate cluster, gene, and isoform number, respectively. Note that in this case a 'gene' refers to a group of transcripts sharing portions of identical sequence. Also provided in the header is the transcript length (299 bases) and the path followed through the assembly graph for transcript construction.*

Identify open reading frames

2. Using TransDecoder, extract all ORFs from the assembled transcripts.

   To do this, type the following command into the terminal:

```
TransDecoder.LongOrfs -t Trinity.fasta -S
```

   Under default settings, TransDecoder will output ORFs of at least 100 amino acids. The `-S` *flag specifies that transcripts are oriented in the sense direction, owing to strand-specific sequencing. Therefore, only the top strand will be examined.*

3. Use TransDecoder to predict the most likely ORFs.

```
TransDecoder.Predict -t Trinity.fasta
```

This step produces several outputs. Among them is a file containing protein sequences of the most likely ORFs, Trinity.fasta.transdecoder.pep.

Create an annotation report

4. Using Trinotate, create and populate a standardized 'boilerplate' database and obtain current releases of Swissprot and Pfam.

```
$TRINOTATE_HOME/admin/Build_Trinotate_Boilerplate_
SQLite_db.pl Trinotate
```

Trinotate relies on protein annotation information from the curated SwissProt and Pfam databases. This step creates the following output:

Trinotate.sqlite (Trinotate boilerplate SQL file)
uniprot_sprot.pep (Swissprot file to be used for BLAST searches)
Pfam-A.hmm.gz (Pfam file to be used for protein domain searches)

5. Create a BLAST protein database from the SwissProt file to enable downstream BLAST analyses.

```
makeblastdb -in uniprot_sprot.pep -dbtype prot
```

6. Unzip the Pfam file and prepare it for use with Hmmscan.

```
gunzip Pfam-A.hmm.gz
hmmpress Pfam-A.hmm
```

7. Conduct a BLASTX search of the Trinity transcripts against the SwissProt database.

```
blastx -query Trinity.fasta -db uniprot_sprot.pep
-num_threads 16 -max_target_seqs 1 -outfmt 6
-evalue 1e-3 -strand plus > trinity.blastx.outfmt6
```

Here, in addition to specifying the query file (Trinity.fasta) and the name of the SwissProt BLAST database, we also include parameters that define the number of threads available on the server (num_threads), the maximum number of hits per query (max_target_seqs), the output format (outfmt), the minimum acceptable E-value (evalue), and that restrict the search to the plus strand only (strand).

8. Similarly, conduct a BLASTP search of the TransDecoder-predicted proteins against the Swissprot database.

```
blastp -query Trinity.fasta.transdecoder.pep -db
uniprot_sprot.pep -num_threads 16 -max_target_seqs 1
-outfmt 6 -evalue 1e-3 >
transdecoder.blastp.outfmt6
```

9. Use Hmmer to identify Pfam protein domains in the TransDecoder-predicted proteins.

```
hmmscan --cpu 16 --domtblout TrinotatePFAM.out Pfam-
A.hmm Trinity.fasta.transdecoder.pep
```

10. Now, load all results into the Trinotate database.

   a. First, load the sequences of transcripts and proteins, as well as the gene-transcript relationships specified in the file `Trinity.fasta.gene_trans_map`. The latter file will be in your Trinity output directory.

```
Trinotate Trinotate.sqlite init --gene_trans_map
Trinity.fasta.gene_trans_map --transcript_fasta
Trinity.fasta --transdecoder_pep Trinity.fasta.
transdecoder.pep
```

   b. Load the BLASTP results.

```
Trinotate Trinotate.sqlite LOAD_swissprot_blastp
transdecoder.blastp.outfmt6
```

   c. Load the BLASTX results.

```
Trinotate Trinotate.sqlite LOAD_swissprot_blastx
trinity.blastx.outfmt6
```

   d. Load the Pfam results.

```
Trinotate Trinotate.sqlite LOAD_pfam TrinotatePFAM.
out
```

11. Finally, output the Trinotate annotation report.

```
Trinotate Trinotate.sqlite report > trinotate_
report.xls
```

   Below is an example of a single transcript annotation from the Trinotate report. It contains the following columns (indicated as C): Trinity gene ID (C1); Trinity transcript ID (C2); top BLASTX hit (C3); protein ID (C4); position of the coding sequence (C5); top BLASTP hit (C6); identified Pfam domain(s) (C7); and associated gene ontology (GO) terms (C8).

```
C1    TRINITY_DN7_c0_g1

C2    TRINITY_DN7_c0_g1_i1

C3    RIP1_PHYAM^RIP1_PHYAM^Q:108-1046,H:1-
      313^100%ID^E:0^RecName: Full=Antiviral protein
      I;^Eukaryota; Viridiplantae; Streptophyta;
      Embryophyta; Tracheophyta; Spermatophyta;
      Magnoliophyta; eudicotyledons; Gunneridae;
      Pentapetalae; Caryophyllales; Phytolaccaceae;
      Phytolacca
```

```
C4      TRINITY_DN7_c0_g1_i1.p1
C5      108-1049[+]
C6      RIP1_PHYAM^RIP1_PHYAM^Q:1-313,H:1-313^100%
        ID^E:0^RecName: Full=Antiviral protein
        I;^Eukaryota; Viridiplantae; Streptophyta;
        Embryophyta; Tracheophyta; Spermatophyta;
        Magnoliophyta; eudicotyledons; Gunneridae;
        Pentapetalae; Caryophyllales; Phytolac-
        caceae; Phytolacca
C7      PF00161.19^RIP^Ribosome inactivating
        protein^35-237^E:6.8e-31
C8      GO:0030598^molecular_function^rRNA
        N-glycosylase activity'GO:0090729^molecular_
        function^toxin activity'GO:0051607^biological_
        process^defense response to virus'GO:
        0017148^biological_process^negative
        regulation of translation
```

MICRORNA PREDICTION AND ANNOTATION

This protocol covers miRNA prediction and the identification of conserved and novel miRNAs. Plant miRNA precursors tend to be longer and more structurally complex than those of animals; therefore, we use miRDeep-P for miRNA prediction since it adopts plant-specific scoring criteria. First, we process the clean small RNA reads to remove undesired species (rRNA, tRNA, snoRNA, snRNA), preventing their mis-annotation as miRNAs. From the filtered small RNA dataset, we prepare a collapsed small RNA fasta file required by miRDeep-P. miRNAs are then predicted as follows: (1) aligning small RNAs to the reference mRNA transcriptome; (2) extracting and folding the mRNA sequence surrounding aligned regions; (3) evaluating small RNAs within the folded structure, based on experimentally determined features of plant miRNA biogenesis. The 'evaluation' step comprises the core algorithm of miRDeep-P. Following prediction, the miRNAs are annotated based on sequence conservation with known plant miRNAs. All input and output files available in the Supporting Information are indicated in code boxes with green and red font, respectively.

Materials

A computer or high-performance cluster running a UNIX-based operating system
Clean small RNA-seq dataset from Basic Protocol 1: `1.sRNA_reads/2.clean`
Reference file of plant non-coding RNA contaminants:
  `6.filter_noncoding_rna/ref_rRNA_tRNA_snoRNA_snRNA.fa`
Perl script used to create a collapsed small RNA fasta file for miRDeep-P:
  `7.mirdeep.p/collapse.pl`
Assembled mRNA transcriptome from Basic Protocol 2:
  `3.trinity/Trinity.fasta`
Predicted miRNA sequences supplied from Neller, Klenov, Guzman, and Hudak
  (2018): `7.mirdeep.p/predicted_miRNAs.fa`
Reference file of known plant miRNAs: `7.mirdeep.p/known_miRNAs.fa`
The following programs will be run in the terminal through the command line.
  They must be installed according to the guidelines at the provided URLs and
  accessible from the user's `$PATH`:
Bowtie (v. 1.2.2): *http://bowtie-bio.sourceforge.net/manual.shtml*

We use Bowtie v.1 rather than v.2 since it is optimized for aligning reads < 50 bp
miRDeep-P (v. 1.3): *https://sourceforge.net/projects/mirdp/*
ViennaRNA (version 2.4.10): *https://www.tbi.univie.ac.at/RNA/*
NCBI BLAST+ (v. 2.7.1): *http://www.ncbi.nlm.nih.gov/books/NBK52640/*

Remove non-coding RNA contaminants

1. Index the contaminant reference file so it can be used with Bowtie.

```
bowtie-build -f --threads 16 ref_rRNA_tRNA_snoRNA_
snRNA.fa ncRNA_junk
```

*Indexing the reference is required for speed and memory efficiency. Here, we spec-
ify that the reference file is in fasta format (`-f`), the number of threads available
(`--threads`), the reference file (`ref_rRNA_tRNA_snoRNA_snRNA.fa`), and a
name for the indexed reference (`ncRNA_junk`).*

*The provided file `ref_rRNA_tRNA_snoRNA_snRNA.fa` contains ~17,000
Viridiplantae non-coding RNA sequences downloaded from the NCBI Nucleotide
database (https://www.ncbi.nlm.nih.gov/nuccore) and the Plant Non-coding RNA
Database (http://structuralbiology.cau.edu.cn/PNRD/).*

2. Using Bowtie, align each clean small RNA file to the indexed reference.

```
bowtie -f -v 3 --norc --un E1_sRNA_junk_filtered.fa
--threads 16 ncRNA_junk sRNA_E1_10k_clean.fasta
```

*We align the small RNA file (`sRNA_E1_10k_clean.fasta`) to the indexed reference
(`ncRNA_junk`) with up to 3 mismatches allowed (`-v 3`). Only the forward strand of the
reference sequence is searched (`--norc`). Importantly, we output the unaligned reads
to a new file (`--un E1_sRNA_junk_filtered.fa`). By retaining only unaligned
reads, this step effectively removes unwanted RNA contaminants that aligned to the junk
file.*

*This step must be repeated for each small RNA input file (six in total), yielding six output
files of unaligned reads.*

Prepare the small RNA file for miRDeep-P

3. Combine the six filtered small RNA files (output from step 2) into a single fasta
file.

```
cat E1_sRNA_junk_filtered.fa E2_sRNA_junk_filtered.fa
E3_sRNA_junk_filtered.fa J1_sRNA_junk_filtered.fa
J2_sRNA_junk_filtered.fa J3_sRNA_junk_filtered.fa >
combined_E_J_sRNA.fa
```

*miRDeep-P takes only one input small RNA file at a time. For ease of use, we create a
combined small RNA file with the `cat` command. This concatenates the files in the order
listed, adding them head-to-tail (i.e., joining the end of file E1 with the start of file E2),
continuing to the end of file J3.*

4. Create a collapsed fasta file from the newly combined small RNA file.

Use the provided script `collapse.pl`.

```
perl collapse.pl combined_E_J_sRNA.fa EJ >
combined_E_J_sRNA_collapsed.fa
```

miRDeep-P requires a collapsed small RNA fasta file, meaning that it contains only non-redundant sequences. The number of identical reads per sequence must be provided in the header. The header has a required format, as shown here:

```
>EJ_0_x66
TGACAGAAGAGAGTGAGCAC
```

EJ is a user-specified prefix appended to every header; 0 is a unique ID number for this read, and x66 indicates the presence of 66 reads having this exact sequence (i.e., summed across the six combined files).

Predict the miRNAs

5. Perform miRNA prediction using miRDeep-P.

   The user must execute several scripts and processing steps prior to the core prediction script. Perl scripts referenced below are in your miRDeep-P installation directory.

   a. Index the reference transcriptome for use with Bowtie.

   ```
   bowtie-build Trinity.fasta Trinity
   ```

   b. Align the collapsed small RNA file to the indexed reference transcriptome using Bowtie.

   ```
   bowtie -f -v 0 -a Trinity combined_E_J_sRNA_
   collapsed.fa >
   sRNA_aligned.sam
   ```

   Only perfect alignments are allowed (-v 0), and alignment information is directed to the output file sRNA_aligned.sam.

   c. Convert the Bowtie output (.sam) to resemble BLAST output format (.blast).

   ```
   perl convert_bowtie_to_blast.pl sRNA_aligned.sam
   combined_E_J_sRNA_collapsed.fa Trinity.fasta >
   sRNA_aligned.blast
   ```

   d. Filter out cases where a small RNA sequence aligned to 15 or more different transcripts.

   ```
   perl filter_alignments.pl sRNA_aligned.blast
   -c 15 >
   sRNA_aligned_15.blast
   ```

   This is a quality control step suggested by the miRDeep-P developers; based on Arabidopsis, a single miRNA family (i.e., duplicated miRNA genes having slight sequence variations) is not expected to comprise more than 15 members. The user can modify this threshold with the -c parameter.

   e. Obtain the sequences of candidate miRNA precursors.

   ```
   perl excise_candidate.pl Trinity.fasta sRNA_aligned_
   15.blast 250 > precursors.fa
   ```

For each small RNA-mRNA alignment, the surrounding 250 nt region of mRNA is excised and sent to a new file, `precursors.fa`.

f. Fold candidate miRNA precursors using the RNAfold program of ViennaRNA.

```
cat precursors.fa | RNAfold > precursors_struct
```

g. Refine alignments by re-aligning small RNAs to the shorter region excised from potential precursors.

> *We issue the same commands as above to index the precursor reference file, align small RNAs, and reformat the output:*

```
bowtie-build -f precursors.fa Precursors
bowtie -f -v 0 -a Precursors combined_E_J_sRNA_
collapsed.fa >
sRNA_precursors_aligned.sam
perl convert_bowtie_to_blast.pl sRNA_precursors_
aligned.sam
combined_E_J_sRNA_collapsed.fa precursors.fa >
sRNA_precursors_aligned.blast
```

h. From the refined alignments produced by the previous step, create a `signa-tures` file.

```
sort +3 -25 sRNA_precursors_aligned.blast >
signatures
```

This step sorts the alignment file, preparing it for the core miRDeep-P prediction script.

i. Run miRDeep-P to obtain miRNA predictions.

```
perl miRDP.pl signatures precursors_struct >
miRNA_predictions
```

Predictions are made based on small RNA-mRNA alignments from the `signatures` *file and folded mRNA structures from the* `precursors_struct` *file. The output file* `miRNA_predictions` *contains sequence, alignment, and structure information for identified miRNAs and corresponding precursors. It also contains scores for each prediction. Top-scoring miRNAs have detectable 'star' reads and precursors that form stable hairpin structures. The star sequence is the passenger strand of the mature miRNA duplex; since it is often degraded, it is usually detected only for highly abundant miRNAs.*

Our sample dataset yielded only two predicted miRNAs. To ensure that there is enough information for downstream analyses, we provide `predicted_miRNAs.fa`. *It contains sequences of the 582 predicted miRNAs obtained by running miRDeep-P on the full datasets in Neller et al. (2018).*

Annotate conserved and novel miRNAs

6. Create a BLAST database of known plant miRNAs.

```
makeblastdb -in known_miRNAs.fa -dbtype nucl
```

The provided reference file of known miRNAs was obtained by downloading all available mature miRNA sequences from the Plant Non-Coding RNA Database (http://structural biology.cau.edu.cn/PNRD/).

7. Conduct a BLASTN search of the predicted miRNAs against known plant miRNAs.

```
blastn -task blastn-short -db known_miRNAs.fa -query
predicted_miRNAs.fa -max_target_seqs 1 -num_threads
16 -outfmt 6 -
strand plus -evalue 1e-3 > blast_annotated_miRNAs.txt
```

We optimize BLASTN for the alignment of short sequences (< 50 nt) with the `-task` *parameter.*

The output file reveals that 23 miRNAs are conserved (i.e., have a significant BLAST hit). Conserved miRNAs tend to have high abundance; for example, a sequence annotated as miR156 has 162,157 counts. All remaining non-annotated miRNAs are considered novel.

BASIC PROTOCOL 4

DIFFERENTIAL EXPRESSION ANALYSIS

This protocol covers mRNA and miRNA differential expression analysis. We begin by quantifying mRNA transcript abundance. The clean RNA-seq reads are aligned back to the reference transcriptome using Bowtie2, then alignments are processed with RSEM to estimate the number of reads derived from each transcript. From these results, we generate a matrix containing the number of reads per transcript for each sample. This count matrix is used as input for differential expression testing with edgeR. miRNA differential expression analysis follows the same procedure, except that quantification with RSEM is not necessary (since one small RNA read = one miRNA). This protocol relies on wrapper scripts packaged with Trinity. Wrapper scripts are helpful for those with limited command-line experience, executing programs 'behind the scene' rather than requiring direct interaction from the user. Note that the programs must be installed correctly and be accessible from the user's $PATH in order to work successfully. All input and output files available in Supporting Information are indicated in code boxes with green and red font, respectively.

Materials

A computer or high-performance cluster running a UNIX-based operating system
Sample information file from Basic Protocol 2: `3.trinity/samples.txt`
Assembled mRNA transcriptome from Basic Protocol 2:
   `3.trinity/Trinity.fasta`
Gene-transcript map from Basic Protocol 2:
   `3.trinity/Trinity.fasta.gene_trans_map`
miRNA count matrix from Neller et al. (2018):
   `8.edgeR/2.DE_miRNA.counts.matrix`
The following programs will be run in the terminal through the command line.
   They must be installed according to the guidelines at the provided URLs and be
   accessible from the user's $PATH:
Trinity (v. 2.8.4): *https://github.com/trinityrnaseq/trinityrnaseq/wiki*
We will use wrapper scripts packaged with Trinity to execute the programs below
Bowtie2 (v. 2.3.4.3): *http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml*
RSEM (v. 1.3.1): *http://deweylab.github.io/RSEM/*
R (v. 3.5.2): *https://www.r-project.org/*
edgeR package (v. 3.8):
   https://bioconductor.org/packages/release/bioc/html/edgeR.html

Prepare mRNA expression matrices

1. For each sample, quantify mRNA transcript abundance with RSEM.

   Use the Trinity wrapper script `align_and_estimate_abundance.pl`:

   ```
   $TRINITY_HOME/util/align_and_estimate_abundance.pl
   --transcripts Trinity.fasta --est_method RSEM
   --aln_method bowtie2 --samples_file samples.txt
   --seqType fq --SS_lib_type RF --thread_count 16
   ```

   This script aligns the clean mRNA-seq reads back to the reference transcriptome using Bowtie2, then processes the alignments with RSEM to estimate the number of reads derived from each transcript.

   Providing `samples.txt` ensures that the outputs for each replicate are organized into a corresponding named directory (i.e., E1, J1, etc.). For each sample, the file `RSEM.isoforms.results` is produced; it contains the estimated number of reads derived from each transcript, as well as a normalized measure of transcript abundance (transcripts per million, TPM) that corrects for transcript length, the number of reads aligned to the transcript, and the total number of reads aligned to all transcripts per sample.

2. Create expression matrices from the individual RSEM results of each sample.

   Use the Trinity wrapper script `abundance_estimates_to_matrix.pl`:

   ```
   $TRINITY_HOME/util/abundance_estimates_to_matrix.pl
   --est_method RSEM --gene_trans_map Trinity.fasta.
   gene_trans_map --
   name_sample_by_basedir./E1/RSEM.isoforms.results./E2/
   RSEM.isoforms.results./E3/RSEM.isoforms.results./J1/
   RSEM.isoforms.results./J2/RSEM.isoforms.results./J3/
   RSEM.isoforms.results
   ```

   This step produces two expression matrices. The file `RSEM.isoform.counts.matrix` contains absolute read counts per transcript; it looks like this:

	E1	E2	E3	J1	J2	J3
TRINITY_DN448_c0_g1_i2	0	0	0	0	0	3
TRINITY_DN2222_c0_g1_i1	0	3	1	2	2	0
TRINITY_DN920_c0_g1_i1	1	1	0	0	1	0

   The file `RSEM.isoform.TMM.EXP.matrix` contains normalized transcript abundance in TPM; it looks like this:

	E1	E2	E3	J1	J2	J3
TRINITY_DN448_c0_g1_i2	0.00	0.00	0.00	0.00	0.00	1328.69
TRINITY_DN2222_c0_g1_i1	0.00	621.56	205.36	441.56	447.10	0.00

```
TRINITY_    194.48 193.31 0.00    0.00    208.21 0.00
DN920_
c0_g1_i1
```

We will use the counts matrix for differential expression testing and the normalized expression matrix to generate a heatmap of transcript abundance.

Identify differentially expressed mRNAs

3. Using the counts matrix from step 2, conduct differential expression testing with edgeR.

   Use the Trinity wrapper script run_DE_analysis.pl:

   ```
   $TRINITY_HOME/Analysis/DifferentialExpression/run_DE_
   analysis.pl --matrix RSEM.isoform.counts.matrix
   --method edgeR --samples_file samples.txt --contrasts
   contrasts.txt
   ```

   This step tests for a significant difference in the read counts of each transcript between two conditions, considering variability present in the biological replicates. The con-trasts.txt file is a tab-delimited text file that specifies the conditions we wish to compare; it contains only this: J(tab)E.

   The output file RSEM.isoform.counts.matrix.J_vs_E.edgeR.DE_re-sults contains results of significance testing, with transcripts ordered from most to least likely to be differentially expressed. Inspecting the first few lines of this file, we see:

   ```
                sampleA   sampleB   logFC   logCPM   PValue    FDR
   TRINITY_     J         E         5.19    15.29    1.21e-96  3.07e-93
   DN1089_
   c0_g1_i1
   TRINITY_     J         E         5.37    15.03    1.64e-93  2.08e-90
   DN797_
   c0_g1_i1
   TRINITY_     J         E         10.81   14.22    5.55e-91  4.67e-88
   DN762_
   c0_g1_i1
   ```

   The following information is provided in each column: transcript ID, sample A name, sample B name, log_2 fold change of sample A counts/sample B counts, average log_2 abundance in counts per million, uncorrected p-value, and false discovery rate (FDR).

   When conducting differential expression at the genome-wide level (i.e., performing a large number of statistical tests), it is important to consider the FDR value, which is a corrected p-value that takes into account multiple hypothesis testing. For example, if we apply a p-value threshold of 0.05, this means that 5% of all tests (i.e., 0.05×3000 transcripts) may be incorrectly detected as being differentially expressed. By contrast, an FDR threshold of 0.05 means that 5% of the significant tests (i.e., $0.05 \times 0.05 \times 3000$) may be incorrectly identified as differentially expressed. Clearly, the FDR value is more stringent than the p-value.

4. Using the normalized expression matrix from step 2, extract and cluster the differentially expressed transcripts.

Use the Trinity wrapper script `analyze_diff_expr.pl`. *This script must be executed from within the edgeR output directory produced in the previous step.*

```
$TRINITY_HOME/Analysis/DifferentialExpression/
analyze_diff_expr.pl --matrix RSEM.isoform.TMM.
EXPR.matrix -P 0.05 -C 0 --samples samples.txt
```

Here, we specify that we wish to extract transcripts having an FDR-corrected p-value (`-P`) cut-off of 0.05. We do not apply a $\log_2$ fold change cut-off (`-C 0`). This step produces several outputs, among which is a clustered heatmap of expression for the 56 identified transcripts meeting these criteria. Note that when working with a full-scale dataset, the user should specify more stringent cut-offs (e.g., `-P 0.001` and `-C 2`, which is 2^2 or 4-fold change).

Identify differentially expressed miRNAs

5. Conduct differential expression testing with edgeR, as shown in step 3.

   Use the provided file `miRNA.counts.matrix`, which contains counts of the 582 predicted miRNAs from Neller et al. (2018).

   ```
   $TRINITY_HOME/Analysis/DifferentialExpression/run_
   DE_analysis.pl --matrix miRNA.counts.matrix --method
   edgeR --contrasts contrasts.txt
   ```

   Note that we do not perform an initial alignment to quantify miRNA abundance as in step 1. This step is not necessary because small RNA-seq captures the miRNA in its entirety (i.e., 1 read = 1 miRNA). The same is not true for mRNA-seq because the transcripts were fragmented during library preparation.

   If readers are using this protocol with their own data, they may need to create the miRNA count matrix. We suggest doing this in Galaxy by counting the number of identical reads in each sample, then joining the independent count tables into a single table. The Galaxy tools 'Group' (invoking the 'Count' operation) and 'Column Join' can be used.

   From the output file `miRNA.counts.matrix.J_vs_E.edgeR.DE_results`, *145 miRNAs are identified as differentially expressed using an FDR cut-off of 0.05. Note that if the reader wishes to create a heatmap of miRNA expression (as in step 4), the counts matrix must be normalized in units of counts (CPM) per million.*

PREDICTION OF microRNA/TARGET PAIRS

In this protocol, we use the plant small RNA analysis server (psRNAtarget) to predict miRNA/target pairs from the subset of differentially expressed miRNAs and mRNAs identified in Basic Protocol 4. We also create a 'background' set for downstream functional enrichment analysis, composed of predicted interactions between all miRNAs and mRNAs. psRNAtarget evaluates sequence complementarity between miRNAs and potential targets using a scoring system benchmarked against experimentally validated plant miRNA/target pairs. All input and output files are available in Supporting Information.

Materials

A personal computer with internet access
The Plant Small RNA Target Analysis Server (psRNAtarget):
   https://plantgrn.noble.org/psRNATarget/home
Sequences of differentially expressed miRNAs identified in Basic Protocol 4:
   `9.psRNAtarget/DE_miRNAs.fa`
Sequences of differentially expressed mRNAs identified in Basic Protocol 4:
   `9.psRNAtarget/DE_mRNAs.fa`

Figure 4 Example of psRNAtarget results. 'miRNA Acc.' and 'Target Acc.' indicate IDs of miRNA and target, respectively; 'Expect' is the score for the miRNA/target pair, with zero indicating perfect sequence complementarity; 'UPE' is the energy required for unpairing secondary structure around the miRNA binding site (not calculated under default parameters); 'Alignment' depicts binding of the miRNA/target pair, with numbers indicating sequence position; 'Target Description' provides additional information in the target ID; 'Inhibition' is a prediction of whether the miRNA/target interaction results in transcript cleavage or translational inhibition (use with caution); 'multiplicity' is the number of miRNA binding sites in the target.

Assembled mRNA transcriptome from Basic Protocol 2:
```
3.trinity/Trinity.fasta
```
Predicted miRNAs from Basic Protocol 3:
```
7.mirdeep.p/predicted_miRNAs.fa
```

1. Navigate to psRNAtarget at: *https://plantgrn.noble.org/psRNATarget/home*.

2. Predict interactions between the differentially expressed miRNAs and mRNAs.

   Under the 'Analysis' tab, select 'Submit small RNAs and targets'. Upload the provided files `DE_miRNAs.fa` and `DE_mRNAs.fa` for the small RNA file and target file, respectively. Run the job with default parameters. Download the results and open them with Microsoft Excel or another spreadsheet program.

   The files `DE_miRNAs.fa` and `DE_mRNAs.fa` contain fasta sequences of the 145 miRNAs and 56 mRNAs identified as differentially expressed in Basic Protocol 4. These sequences were extracted from `predicted_miRNAs.fa` and `Trinity.fa`. Readers can prepare analogous files for their own data using the `seqtk_subseq` tool in Galaxy.

   In this example, 89 miRNA/target interactions were predicted, comprising 43 unique targets. This means that some mRNAs were predicted to be targeted by more than one miRNA. For the purpose of this example, we did not apply any filters on these predictions. Readers may wish to lower the Expectation cut-off (e.g., from 5 to 3) to identify the highest-confidence predictions when working with their own full-scale dataset.

3. Repeat step 2 to predict interactions between all miRNAs and mRNAs.

   Upload `predicted_miRNAs.fa` and `Trinity.fa` for the small RNA file and target file, respectively.

   An example of psRNAtarget results is shown in Figure 4. These predicted interactions will form the background set for functional enrichment analysis in Basic Protocol 6.

FUNCTIONAL ENRICHMENT ANALYSIS OF DIFFERENTIALLY EXPRESSED microRNA/TARGET PAIRS

This protocol covers GO term functional enrichment analysis. GO terms are structured, curated concepts relating to gene functions. A gene product may be annotated in up to three categories: molecular function, cellular component, and biological process. The ontology is structured as a directed graph in which each term has defined relationships to one or more other terms. Lower-level terms have greater specificity than higher-level ones. In this protocol, we use GOseq to identify functional associations that are over-represented among the differentially expressed miRNA/target pairs. GOseq also corrects for transcript length bias: the fact that longer transcripts are more likely to be identified as differentially expressed than shorter transcripts. Note that GO terms are only ascribed to proteins; therefore, this analysis is conducted on targets of miRNAs, not the miRNAs themselves. All input and output files available in Supporting Information are indicated in code boxes with green and red font, respectively.

Materials

A computer or high-performance cluster running a UNIX-based operating system

Trinotate report from Basic Protocol 2: `5.trinotate/trinotate_report.xls`

Assembled mRNA transcriptome from Basic Protocol 2: `3.trinity/Trinity.fasta`

Provided target ID file for the test set (differentially expressed miRNA/target pairs) identified in Basic Protocol 5, step 2: `10.goseq/DE_targets_of_DE_miRNAs.txt`

Provided target ID file for the background set (all miRNA/target pairs) identified in Basic Protocol 5, step 3: `10.goseq/targets_of_all_miRNAs.txt`

The following programs will be run in the terminal through the command line. They must be installed according to the guidelines at the provided URLs and accessible from the user's `$PATH`:

Trinotate (v. 3.1.1): *https://github.com/Trinotate/Trinotate.github.io/wiki*

Trinity (v. 2.8.4): *https://github.com/trinityrnaseq/trinityrnaseq/wiki*

We will use a wrapper script packaged with Trinity to execute the programs below:

R (v. 3.5.2): *https://www.r-project.org/*

GOseq package (v. 3.8): *https://bioconductor.org/packages/release/bioc/html/goseq.html*

1. Extract all GO terms from the Trinotate annotation report.

```
$TRINOTATE_HOME/util/extract_GO_assignments_from_
Trinotate_xls.pl --Trinotate_xls trinotate_report.
xls --trans --include_ancestral_terms >go_
annotations.txt
```

The parameter 'include ancestral terms' means that for each GO term, all terms 'upstream' in the GO hierarchy will also be extracted. This is important to account for differences in annotation specificity/depth between transcripts.

2. Create a transcript lengths file for use with GOseq.

```
$TRINITY_HOME/util/misc/fasta_seq_length.pl Trinity.
fasta >
Trinity.fasta.seq_lens
```

3. Run GOseq to test for enriched GO terms associated with differentially expressed miRNA/target pairs.

Provide the extracted GO terms from step 1, the lengths file from step 2, and the mRNA transcript IDs of targets in the test set (differentially expressed miRNA/target pairs) and reference set (all miRNA/target pairs).

```
$TRINITY_HOME/Analysis/DifferentialExpression/run_
GOseq.pl --genes_single_factor DE_targets_of_DE_
miRNAs.txt --GO_assignments go_annotations.txt
--lengths Trinity.fasta.seq_lens --background
targets_of_all_miRNAs.txt
```

This step tests for significant enrichment of GO terms in the test set relative to the background set. The parameters 'genes single factor' and 'background' specify target IDs of the test set and background set, respectively. Readers can produce target ID files for their own data by copying the 'Target_Acc' (i.e., target accession) column from each psRNAtarget output of Basic Protocol 5, pasting it into a text editor like Notepad, and saving the file.

The output file DE_targets_of_DE_miRNAs.txt.GOseq.enriched *contains GO terms that are significantly enriched in the test set (p < 0.05). Results are ordered from most to least significantly enriched, with both the p-value and FDR value provided.*

From our sample data, 32 GO terms are significantly enriched (p < 0.05), but none are enriched at the FDR cut-off (FDR < 0.05). Since we are working on a relatively small example dataset, the FDR value may be too stringently applied here. Nevertheless, inspecting the top few enriched GO terms in the list provides insight into the biological relevance of interactions among differentially expressed miRNA/target pairs: NADPH activity; toxin activity; carbonate dehydratase activity; coenzyme A metabolic process; nucleoside bisphosphate metabolic process; ribonucleoside bisphosphate metabolic process; purine nucleoside bisphosphate metabolic process; defense response. Since JA is known to mediate plant defense against pathogens, we may wish to investigate targets annotated with the GO term 'defense response'. The enrichment results file reveals that five targets in the test set are annotated with this term. We can look up their IDs in the Trinotate report to obtain their top BLAST-annotated hits, if available, for which four of five are: Antiviral protein 1; Antiviral protein 2; 2-Cys peroxiredoxin BAS1-like; Polygalacturonase inhibitor. Having identified some interesting miRNA/target interactions, the next step is to validate them.

DETECTION OF microRNA-INDUCED CLEAVAGE USING RNA LIGASE–MEDIATED 5′ RACE

In this protocol, we cover the validation of an miRNA/target pair using RNA ligase–mediated 5′ RACE (RLM-RACE; Fig. 5). Specifically, the 5′ end of the expected cleaved target is cloned and sequenced to assess whether it underwent slicing by a miRNA-guided Argonaute protein. This method takes advantage of the exposed 5′ monophosphate on the target that results from miRNA-induced cleavage. First, the poly(A) mRNA pool (from the Support Protocol) undergoes 5′ ligation with an RNA adapter. Then, the mRNA pool is reverse transcribed into cDNA using a 3′ d(T) adapter. To enrich for low-abundance cleavage products, the cDNA pool is subjected to nonspecific PCR amplification using primers annealing to the RNA adapter and 3′ d(T) adapter. Finally, the specific target is amplified using a 5′ adapter primer and 3′ gene-specific primer. Sequences to facilitate cloning into a vector using Gibson assembly are introduced by a subsequent PCR of the specific target. The cloned target is transformed into *E. coli*, with successful transformants identified by colony PCR, and plasmid DNA is sequenced. An miRNA/target pair is considered validated if, after alignment, the 5′ end of the cloned sequences lines up with the expected miRNA cut site on the mRNA.
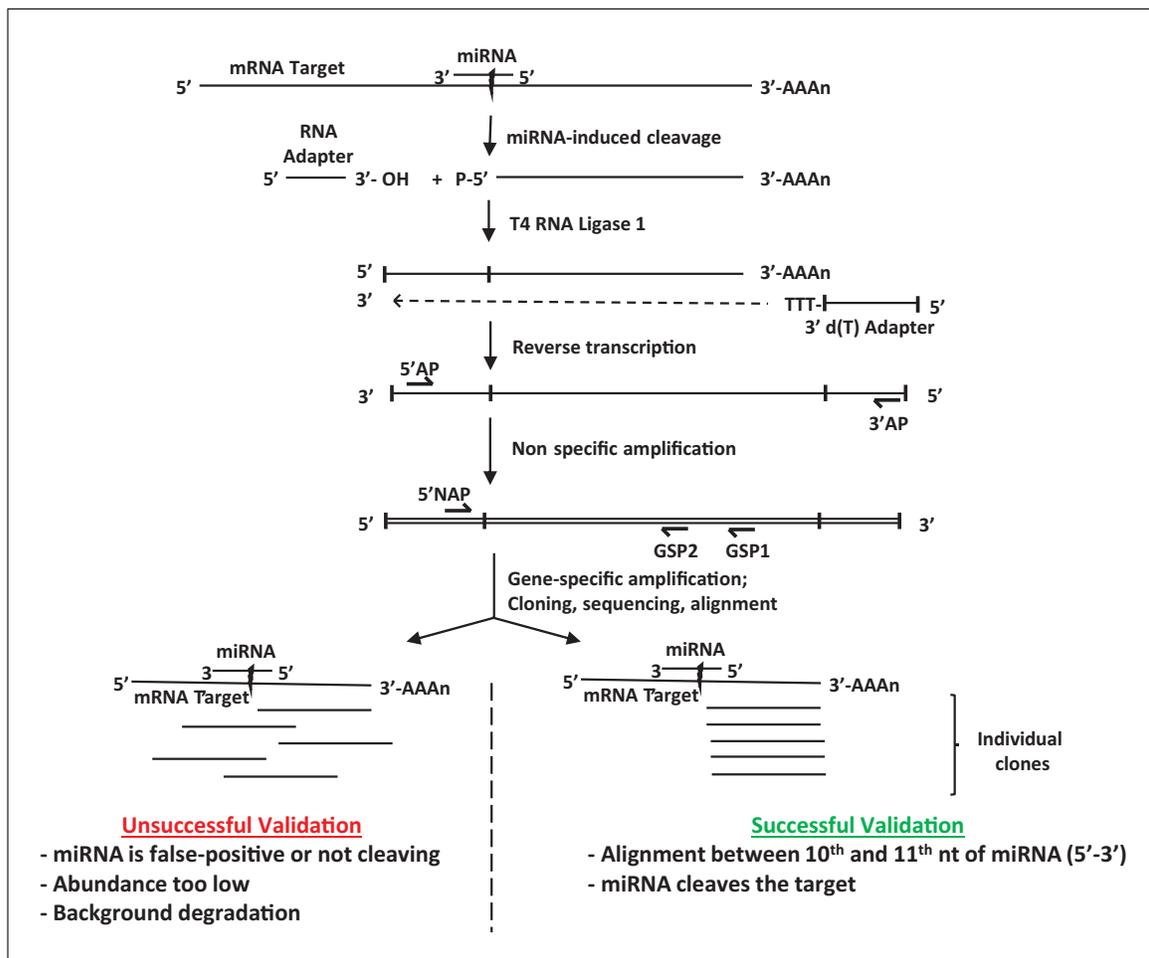
Figure 5 Overview of miRNA cleavage detection by modified RLM-RACE. miRNAs that induce cleavage leave exposed 5′ phosphates on their target mRNAs. This allows ligation of an RNA adapter, followed by nonspecific PCR amplification with 5′ and 3′ adapter primers (5′AP, 3′AP) and gene-specific amplification with 5′ nested adapter primers (5′ NAP) and gene-specific primers (GSP1, GSP2). PCR products of the expected size are cloned, sequenced, and aligned to the target mRNA sequence. Sequences that align between the 10th and 11th nucleotide of the miRNA binding site indicate successful validation.

While kits for RLM-RACE are available, they contain reagents not relevant here, which select for full-length, capped mRNAs. Therefore, we did not purchase any kit, though our protocol is based on the GeneRacer RLM-RACE kit (ThermoFisher L1500-01). The RNA adapter and all primers were synthesized separately, and decapping and dephosphorylating steps were omitted.

Materials

Poly(A) mRNA pool (Support Protocol)
Primers (and RNA adapters) in Table 1
10× T4 RNA ligase buffer (NEB B0216L)
10 mM ATP (NEB P0756S)
40 U/μl murine RNase inhibitor (NEB M0314S)
10 U/μl T4 RNA ligase 1 (NEB M0204S)
DOW Corning High Vacuum Grease, autoclaved in 5-ml syringe without needle
Acid PCI: 25:24:1 (v/v/v) citrate-buffered phenol/chloroform/isoamyl alcohol
  [25 ml citrate-buffered phenol (Sigma P4682), 24 ml chloroform, 1 ml isoamyl alcohol]
Precipitation mix: 50 ml of ethanol, 2 ml of 3 M sodium acetate, pH 5.2
5 mg/ml linear acrylamide (ThermoFisher AM9520)

10 mM dNTPs (NEB N0446S)
5× first-strand buffer (see recipe)
0.1 M DTT
200 U/μl Superscript III Reverse Transcriptase (ThermoFisher 18080093)
RNase H (NEB M0297S)
5× Q5 polymerase reaction buffer (NEB B9027S)
2 U/μl Q5 DNA polymerase (NEB M0491S)
EZ-10 Spin Column Plasmid DNA Miniprep Kit (BioBasic BS413)
EZ-10 Spin Column DNA Gel Extraction Kit (Biobasic BS353)
Cloning vector (pHSG298; Takara 3298)
2× Gibson assembly mastermix (NEB E2611S)
5-α competent *E. coli* (NEB C2987I)
LB plates with appropriate antibiotic for cloning vector
2 U/μl pfu-sso7d or equivalent polymerase
5× pfu-sso7d buffer (see recipe)

Thermal cycler with gradient function
UV spectrometer
0.2-ml PCR tubes
Sanger sequencing facility

Additional reagents and equipment for agarose gel electrophoresis (Voytas, 2000)

Ligate the 5′ RNA adapter to the mRNA pool

1. Combine 250 ng of poly(A) mRNA and 250 ng of RNA adapter (sequence provided in Table 1) in a total volume of 7 μl. Incubate at 65°C for 5 min to denature. Chill on ice for 2 min.

   The GeneRacer manual suggests using 100 to 250 ng of poly(A) mRNA as input. We recommend beginning with at least 250 ng due to losses from multiple precipitation steps.

2. Add the following reagents to the denatured mRNA:

   1 μl 10× T4 RNA ligase buffer
   1 μl 10 mM ATP
   1 μl 40 U/μl murine RNase inhibitor
   0.5 μl 10 U/μl T4 RNA ligase 1
   0.5 μl distilled H$_2$O.

3. Incubate at 37°C for 1 hr to ligate, then place on ice.

4. Into a new 1.5-ml microcentrifuge tube, dispense 50 μl of autoclaved high vacuum grease using a syringe. Microcentrifuge for 30 sec at 16,000 × *g*, 4°C, to collect grease at the bottom of the tube. Add the ligated mRNA, 90 μl distilled H$_2$O, and 100 μl of acid PCI. Vortex for 30 sec, then microcentrifuge 5 min at 16,000 × g, 4°C.

   The vacuum grease provides a physical barrier between the organic phase (containing phenol, chloroform, and protein) and the aqueous phase (containing RNA). This increases RNA yield and purity.

5. Transfer the supernatant to a new 1.5-ml tube. Add 2.6 volumes of precipitation mix and 4 μl of 5 mg/ml linear acrylamide. Precipitate on dry ice for 10 min or at −20°C overnight.

   A precipitant is recommended due to the low concentration of RNA; we use linear acrylamide here. Derived from a synthetic source, it is a superior alternative to glycogen or salmon sperm DNA, which can introduce foreign nucleic acids into the reaction.

Table 1 Primers for RLM-RACE

Primer name	Sequence (5′ to 3′)	Usage
RNA Adapter	CGACUGGAGCACGAGGACACUG ACAUGGACUGAAGGAGUAGAAA	Basic Protocol 7, step 1
3′ d(T) Adapter	GCTGTCAACGATACGCTACGTAAC GGCATGACAGTG-d(T)$_{24}$	Basic Protocol 7, step 6
3′ Adapter Primer (3′AP)	GCTGTCAACGATACGCTACGTAACG	Basic Protocol 7, step 11
5′ Adapter Primer (5′AP)	CGACTGGAGCACGAGGACACTGA	Basic Protocol 7, step 11, 13
Gene Specific Primer 1 (GSP1)	. . . XXXXXXXX . . .	Basic Protocol 7, step 13
5′ Nested Adapter Primer (5′NAP)	GGACACTGACATGGACTGAAGG AGTA	Basic Protocol 7, step 17
Gene Specific Primer 2 (GSP2)	. . . XXXXXXXX . . .	Basic Protocol 7, step 16
5′ Cloning Nested Adapter Primer	. . . NNNNNNNN-GGACACTG ACATGGACTGAAGGAGTA	Basic Protocol 7, step 20
Cloning Gene Specific Primer 2	. . . NNNNNNNN-XXXXXXXX . . .	Basic Protocol 7, step 20
T7 Promoter Primer	TAATACGACTCACTATAGGG	Basic Protocol 7, step 26
T7 Terminator Primer	GCTAGTTATTGCTCAGCGG	Basic Protocol 7, step 26

[a]N's indicate nucleotides added to the primer to facilitate Gibson assembly, and will share ~15 to 30 bp of sequence complementarity with the cloning vector. X's indicate sequence complementarity with the mRNA under validation.

Centrifuge the ligated mRNA 20 min at 16,000 × g, 4°C. Discard the supernatant and add 1 ml of 75% ethanol. Invert the tube several times, then centrifuge at 16,000 × g, 4°C for an additional minute. Pour off the supernatant, quick-spin once more with the same conditions, and pipette off the remaining supernatant. Resuspend the pellet in 11 µl distilled H$_2$O.

Reverse transcribe the ligated mRNA pool

6. Add the following reagents to the mRNA:

    1 µl 10 µM 3′ d(T) adapter
    1 µl 10 mM dNTPs.

7. Denature the mRNA at 65°C for 10 min. Place on ice for at least 2 min. Add the following reagents and mix well:

    4 µl 5× first-strand buffer
    1 µl 0.1 M DTT
    1 µl 40 U/µl murine RNase inhibitor
    1 µl 200 U/µl Superscript III Reverse Transcriptase.

8. Incubate at 50°C for 1 hr to reverse transcribe the mRNA into cDNA.

9. Incubate at 70°C for 15 min to inactivate the reverse transcriptase.

Neller et al.
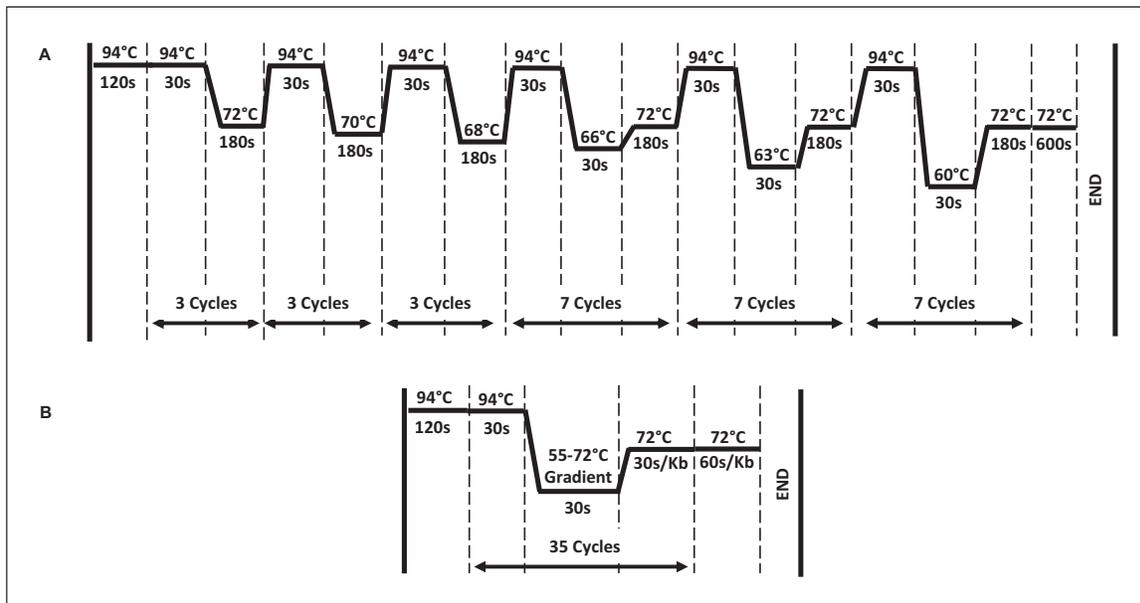
Figure 6 Thermal cycling profiles for RLM-RACE. (**A**) Touchdown PCR is performed for nonspecific amplification of the cDNA pool. (**B**) For all other PCR amplifications, a standard profile is used, with or without a gradient annealing step. The number of cycles is indicated, along with respective temperature (°C) and duration (seconds).

10. Add 2 U of RNase H to the reverse transcription reaction and incubate at 37°C for 20 min.

   Store cDNA at −20°C until needed.

Amplify the cDNA pool

11. Using 1 μl of reverse transcription product as template, set up the following reaction to perform nonspecific PCR amplification of the cDNA pool:

   5 μl 5× Q5 polymerase buffer
   1.25 μl 10 μM 5′ adapter primer
   1.25 μl 10 μM 3′ adapter primer
   0.5 μl 10 mM dNTPs
   1 μl reverse transcription product as template
   0.25 μl 2 U/μl Q5 DNA polymerase
   15.75 μl distilled H_2O.

12. Place into a thermal cycler and amplify with the touchdown program shown in Figure 6A.

   Nonspecific amplification of the cDNA pool increases the chance of detecting cleavage for low-abundance miRNAs and/or targets.

Amplify the specific cDNA target

13. Using 1 μl of nonspecific PCR product as template, set up the following reaction for the first gene-specific amplification:

   5 μl 5× Q5 polymerase buffer
   1.25 μl 10 μM 5′ adapter primer
   1.25 μl 10 μM gene specific primer 1
   0.5 μl 10 mM dNTPs
   1 μl nonspecific PCR product as template
   0.25 μl 2 U/μl Q5 DNA polymerase
   15.75 μl distilled H_2O.

To validate different mRNA targets simultaneously, multiple reactions may be prepared, each containing a target-specific reverse primer.

14. Place into a thermal cycler and amplify as shown in Figure 6B.

    A gradient thermocycler is not essential, but its use can save significant time on PCR optimization.

15. Separate 8 µl of gene-specific PCR product on a 1.2% agarose gel, then visualize.

    Agarose gel electrophoresis is described in Voytas (2000). A smear, multiple bands, or lack of visible product may be seen at this stage, depending on abundance of cleaved target and primer specificity.

    To determine the anticipated product size, take into account the length of the 5′ forward adapter primer and distance between the cleavage site and gene-specific reverse primer.

    If a discrete band of the expected size is present, proceed directly to step 19 for gel purification. Otherwise, continue with step 16 for nested PCR.

16. Dilute the gene-specific PCR product 1:10 with distilled H_2O and use 1 µl as template in the following reaction to perform nested PCR. Use the program in Figure 6B.

       5 µl 5× Q5 polymerase buffer
       1.25 µl 10 µM 5′ nested adapter primer
       1.25 µl 10 µM gene specific primer 2
       0.5 µl 10 mM dNTPs
       1 µl diluted gene-specific PCR product
       0.25 µl 2 U/µl Q5 DNA polymerase
       15.75 µl distilled H_2O.

17. Separate the PCR product on a gel as in step 15.

18. If a single band is observed at the expected size, dilute the reaction 1:10 with distilled H_2O and proceed to amplification, step 21, for cloning. If multiple bands are observed, separate the remainder of the PCR reaction on a 1% low-melt agarose gel (Voytas, 2000) and excise the band corresponding to the expected product size.

    Consider using a nucleic acid binding dye that emits green light under blue light excitation (e.g., SYBR Green, Midori Green) to reduce exposure of DNA to UV light that is used to visualize classic ethidium bromide staining.

19. Purify the gel fragment with the EZ-10 Spin Column DNA Gel Extraction Kit according to the manufacturer's protocol, then quantify with a UV spectrometer.

    To maximize recovery of gel-purified DNA, run the melted gel fragment through the silica column twice, elute twice. each time with 50 µl elution buffer, then concentrate with a vacuum concentrator to a final volume of 20 µl.

Clone the amplified cDNA target using Gibson assembly

20. Use 1 µl of diluted PCR product from step 18 or 1 ng of purified fragment from step 19, and set up the following reaction. Use the PCR program in Figure 6B:

       5 µl 5× Q5 polymerase buffer
       1.25 µl 10 µM 5′ cloning nested adapter primer
       1.25 µl 10 µM cloning gene specific primer 2
       0.5 µl 10 mM dNTPs
       1 µl diluted PCR or gel-extracted product (from step 18 or 19)
       0.25 µl 2 U/µl Q5 DNA polymerase
       15.75 µl distilled H_2O.

A separate PCR reaction enables addition of overhangs for Gibson assembly, using the cloning primers in step 21.

21. Gel purify (steps 18 and 19) the resultant PCR product of the expected size and quantify with a UV spectrometer.

22. Combine the cloning vector pHSG298 and insert (i.e., the purified product from step 21) at a 1:2 molar ratio. Also perform a 1:0 reaction as a negative control. Perform Gibson assembly according to manufacturer's protocol.

    Ideally, use 50 fmol of vector and 100 fmol of insert, although attempts with 20 fmol vector have been successful. Inserts of less than 250 bp may require a vector to insert ratio of 1:5.

23. Transform 1 to 5 µl of Gibson assembly product into 50 µl of competent NEB 5-α *E. coli,* per manufacturer's instructions. Plate onto LB plates with appropriate antibiotics and incubate at 37°C overnight.

Screen transformants using colony PCR

24. Select a 1-mm-diameter colony from the experimental plate and resuspend in 50 µl distilled H_2O. Spot 1 µl of the suspension onto an antibiotic patch plate to preserve the clone.

    A patch plate is prepared by drawing a grid on the bottom of a petri dish containing medium and antibiotic, and labelling each region to identify the spotted suspension.

    We recommend screening 10 clones per target.

25. Lyse the resuspended *E. coli* by incubating at 95°C for 5 min. Centrifuge 10 min at $16,000 \times g$, room temperature, to pellet cell debris. Transfer 10 µl of supernatant to a fresh 0.2-ml PCR tube.

    A white pellet can be seen if enough E. coli has been lysed for a successful colony PCR.

26. Set up the following reaction to perform colony PCR. Use primers that flank the cloning site in your vector, typically the T7 promotor and terminator.

    5 µl 5× pfu-sso7d buffer
    1.25 µl 10 mM T7 promotor primer
    1.25 µl 10 mM T7 terminator primer
    0.5 µl 10 mM dNTPs
    10 µl bacterial lysate (from step 26)
    0.25 µl 2 U/µl pfu-sso7d polymerase
    15.75 µl distilled H_2O.

    Since many colonies are screened and the template DNA is generally of low complexity, we use a more economical polymerase here rather than a high-end enzyme like Q5.

    Amplify with the PCR program illustrated in Figure 6B, using 58°C rather than the gradient annealing step.

27. Separate colony PCR products on a 1% agarose gel, then visualize. For colonies containing the appropriately sized insert, grow liquid cultures using the patch plate from step 24. Extract plasmid DNA from these cultures using a BioBasic EZ-10 Spin Column Plasmid DNA Miniprep Kit and send to a facility for Sanger sequencing with a forward and reverse primer flanking the cloning site.

    Agarose gel electrophoresis is described in Voytas (2000). We recommend sequencing 5 to 10 clones per target.
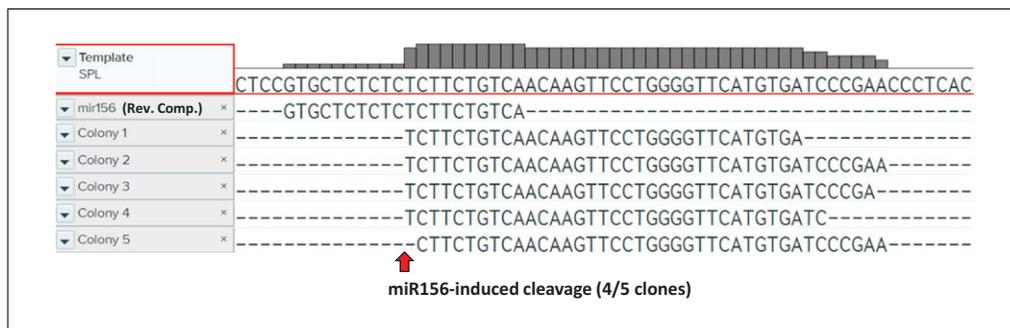
Figure 7 Validation of SPL mRNA cleavage by miR156. The RLM-RACE protocol was performed for the positive control miRNA/target pair, miR156/SPL. The miR156 sequence was reverse-complemented by the MAAFT aligner for easy visualization of miRNA and target alignment. Sequences of five individual clones of cleaved SPL mRNA are shown, with 4/5 indicating the expected cleavage site.

Analyze the alignments

28. Upon obtaining the sequences, align them to the target mRNA sequence using software. We recommend the online tool Benchling (*http://www.benchling.com*).

   a. Create an account and log on to Benchling. In the tool bar on the left, select the 'Projects' tab. Click on the (+) icon to import your mRNA sequence.
   b. In the tool bar on the right, click on 'Alignments' and select 'Create New Alignment'. Input the sequences of your individual clones as well as the sequence of the putative miRNA. Choose MAFFT as your alignment algorithm. Click 'Create Alignment' and view your results. If your positive control is miR156 with target SPL, the alignment of your colony sequences should resemble Figure 7.

   miRNAs induce cleavage between the 10th and 11th nucleotides, 5' to 3', of the miRNA. 5' ends of cloned sequences that align to this site indicate successful validation of cleavage for the miRNA/target pair.

   Cleavage at unexpected sites suggests that the miRNA is not cleaving the target as anticipated, or that this small RNA is not a functional miRNA (i.e., false positive). In some cases, we observed randomly distributed cleavage surrounding the expected site. This suggested that the limit of detection had been reached: background levels of mRNA degradation were as high as the miRNA cleavage. Thus, it is difficult to obtain trustworthy, reproducible results with RLM-RACE when abundance of the miRNA and/or target is low (e.g., < 10 counts per million and < 100 transcripts per million respectively).

TOTAL RNA EXTRACTION AND ISOLATION OF SMALL RNA AND mRNA

SUPPORT PROTOCOL

In this protocol, we cover total RNA extraction from plant tissue and steps to enrich for either small RNA or mRNA. RNAzol reagent is used to enable extraction of both total RNA and small RNA from a single plant sample. Total RNA is obtained first. Plant tissue is ground in liquid nitrogen, followed by addition of RNAzol; DNA and protein contaminants are removed through centrifugation, then total RNA is precipitated with ethanol. Using the supernatant reserved from total RNA extraction, the small RNA fraction is then obtained by isopropanol precipitation. mRNA enrichment is performed using the NEBNext Poly(A) mRNA Magnetic Isolation Module on the previously extracted total RNA. Starting with high-quality RNA is critical for the success of downstream validations. Specifically, total RNA, small RNA, and mRNA will be used for qRT-PCR and stem-loop qRT-PCR (Basic Protocol 8), and RLM-RACE (Basic Protocol 7), respectively. Small-RNA enrichment serves to concentrate low-abundance miRNAs to enable their detection. Similarly, mRNA enrichment concentrates low-abundance targets; it also removes non-polyadenylated transcripts and ribosomal RNA so they will not be cloned inadvertently during RLM-RACE.

Neller et al.

Materials

    Fresh leaf tissue of interest
    RNAzol reagent (Molecular Research Center RN190)
    100% and 75% (v/v) ethanol

    RNA storage buffer [1 mM citrate buffer, pH 6.4 (see recipe) containing 22.5 mM
       DTT]
    100% isopropanol
    NEBNext Poly(A) mRNA magnetic isolation module (NEB E7490S)

Extract total RNA

1. Grind 100 mg of fresh leaf tissue to a fine powder with liquid nitrogen in a 1.5-ml microcentrifuge tube. Add 1 ml RNAzol directly to the tissue and immediately vortex to homogenize.

2. Add 0.4 ml distilled H_2O to the RNAzol and vortex briefly. Incubate at room temperature for 10 min.

3. Microcentrifuge the lysate 15 min at $16,000 \times g$, room temperature, to pellet DNA and protein. Carefully remove 1 ml of the supernatant (75%) and transfer it to a clean 1.5-ml tube.

   The pellet tends to expand quickly; work with small batches of 4 to 8 tubes at a time while centrifuging the remainder.

4. Add 0.4 ml 100% ethanol to the tube and place on ice for 10 min. Microcentrifuge 20 min at $16,000 \times g$, 4°C, to pellet total RNA. Transfer supernatant to a fresh tube and keep on ice for step 7.

   The pellet contains total RNA >200 bp in length. The supernatant contains small RNA, and is retained for isolation in step 7.

5. Add 1 ml 75% ethanol to the total RNA pellet. Invert the tube several times, then microcentrifuge 1 min at $16,000 \times g$, 4°C. Pour off the ethanol and microcentrifuge under the same conditions for 30 sec. Pipette off remaining liquid and discard.

6. Resuspend the total RNA pellet in 20 µl of pre-warmed (50°C) RNA storage buffer. Quantify the RNA with UV spectroscopy by measuring the absorbance at 260 nm.

   One A_{260} unit of single-stranded RNA = 40 µg/ml. The quality of RNA can be assessed by measuring absorbance at 260/280 nm. A ratio of 2.0 is generally accepted as good-quality RNA.

   Store total RNA at −80°C until needed.

Isolate small RNA

7. To the saved supernatant from step 4, add 0.8 volumes 100% isopropanol. Chill on ice for 30 min. Centrifuge 20 min at $16,000 \times g$, 4°C, to pellet small RNA. Pour off the supernatant.

8. Repeat steps 5 and 6 to purify and resuspend the small RNA.

   Store small RNA at −80°C until needed.

Isolate mRNA

9. Using 30 µg of total RNA from step 6, carry out mRNA enrichment according to the product manual of the NEBNext Poly(A) mRNA magnetic isolation module. Upon

completion, quantify the RNA with UV spectroscopy by measuring the absorbance at 260 nm.

One A_{260} unit of single-stranded RNA = 40 μg/ml. The quality of RNA can be assessed by measuring absorbance at 260/280 nm. A ratio of 2.0 is generally accepted as good-quality RNA.

To isolate enough mRNA for downstream analysis and reduce the number of samples to be processed, we suggest using 50 μl of magnetic beads instead of the 20 μl stated by the manufacturer.

Store mRNA at −80°C until needed.

One milligram of leaf tissue will yield 0.5 to 1.5 μg of total RNA, 50 to 150 ng of small RNA, and 15 to 50 ng of mRNA. Average yield is heavily dependent on plant species, with high polysaccharide content being particularly inhibitory.

QUANTIFICATION OF microRNA AND TARGET EXPRESSION USING qRT-PCR

Here, we cover validation of miRNA and target differential expression results from RNA-seq using stem-loop qRT-PCR and standard qRT-PCR, respectively. Both methods rely on the same underlying principle: a relative fold change is determined by quantifying RNA levels of the species of interest under each condition and normalizing the expression change to an internal control that is stably expressed for the treatment under study. To prepare for qPCR, the RNA of interest and the internal control are both reverse transcribed into cDNA using gene-specific primers. Then, qPCR is conducted to quantify cDNA levels, based on fluorescence occurring from intercalation of dye into the double-stranded PCR product. In terms of primer design for stem-loop and standard qRT-PCR (Fig. 8), the latter is straightforward in that primers must simply produce a product of 150 to 200 bp. In contrast, the short length of miRNAs (~21 nt) necessitates use of a specialized stem-loop primer during reverse transcription to increase the melting point
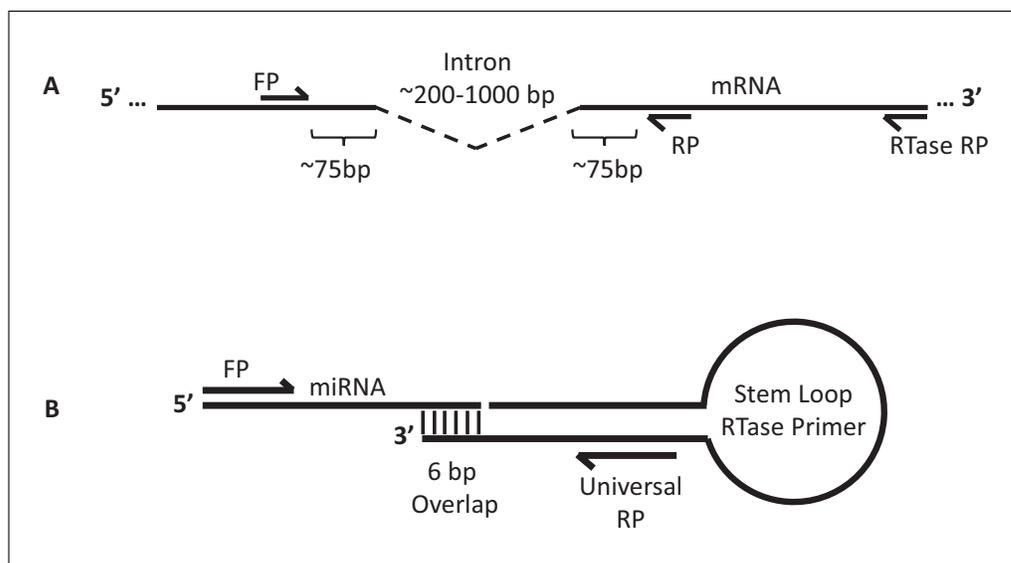
Figure 8 Primer design for qRT-PCR validation of mRNA and miRNA expression. (**A**) The ideal product size for a qRT-PCR reaction is between 150 and 200 bp. Wherever possible, forward primers (FPs) and reverse primers (RPs) should be designed to span an intron, which will reduce the level of genomic DNA amplification from contamination and allow its detection on an agarose gel for troubleshooting. (**B**) miRNAs require reverse transcription with a reverse primer bearing a stem-loop, which serves to increase the length and melting point of the PCR product, to be compatible with standard PCR cycling.

Table 2 Primers for qRT-PCR

Primer name	Sequence (5′ to 3′)	Usage
mRNA/Internal Control Reverse Transcriptase Primer (RTase RP)	. . . XXXXXXXX . . .	Basic Protocol 8, step 1
miRNA/Internal Control Stem Loop RTase Primer	GTCGTATCCAGTGCAGGGTCCGA GGTATTCGCACTGGATACGAC-XX XXXX	Basic Protocol 8, step 1
mRNA/Internal Control Forward Primer (FP)	. . . XXXXXXXX . . .	Basic Protocol 8, step 4
mRNA/Internal Control Reverse Primer (RP)	. . . XXXXXXXX . . .	Basic Protocol 8, step 4
miRNA/Internal Control Forward Primer	. . . XXXXXXXX	Basic Protocol 8, step 4
Universal Reverse Primer (RP)	CCAGTGCAGGGTCCGAGGTA	Basic Protocol 8, step 4

[a]X's indicate sequence complementarity with the miRNA/mRNA under validation.

of the product, followed by PCR using a reverse primer complementary to the stem-loop and a forward primer specific to the miRNA.

Following qPCR, the $\Delta\Delta$Ct formula is used to determine the relative fold change. The Ct value is the PCR cycle at which fluorescence rises above a set threshold during exponential amplification. The lower the Ct value, the higher the abundance of the RNA under study. A statistical test is conducted to determine whether the RNA abundance is significantly different between the experimental and control conditions. Finally, the fold changes obtained by qRT-PCR are correlated with those obtained by RNA-seq to determine whether the two methods of RNA quantification agree. Note that although most miRNA/target pairs under validation should be randomly selected, those of special interest are also tested.

Materials

Total RNA or small RNA of interest (Support Protocol)
Primers in Table 2
10 mM dNTPs (NEB N0446S)
5× first-strand buffer (see recipe)
0.1 M DTT
40 U/μl murine RNase inhibitor (NEB M0314S)
200 U/μl Superscript III reverse transcriptase (ThermoFisher 18080093)
2× SYBR Green mastermix (Bimake B21202)

0.2-ml PCR tubes
Rotor-GeneQ qRT-PCR System (Qiagen)

Reverse transcribe the total RNA and small RNA

1. Dilute 250 ng of total RNA or small RNA (Support Protocol) in 11 μl distilled H_2O. Add the following reagents:

   1 μl 10 μM mRNA reverse transcriptase primer *or* stem-loop RTase primer
   1 μl 10 mM dNTPs.

Prepare separate samples for each miRNA and target. For miRNA samples, use the stem-loop reverse primer shown in Table 2; for target samples, use a standard reverse primer.

Prepare the appropriate internal controls for both the miRNA and mRNA. Internal controls are chosen based on their levels remaining stable for the treatment under study. Guidance in choosing appropriate internal controls is included in the Critical Parameters section. At least two miRNA and two mRNA internal controls should be included.

2. Denature the RNA at 65°C for 10 min, then place on ice for at least 2 min. Add the following reagents:

   4 μl 5× first-strand buffer
   1 μl 0.1 M DTT
   1 μl 40 U/μl murine RNase inhibitor
   1 μl 200 U/μl Superscript III reverse transcriptase.

3. Mix well, then incubate at 50°C for 1 hr to perform reverse transcription. Inactivate the reverse transcriptase at 70°C for 15 min.

   Store cDNA at −20°C until needed.

Quantify miRNA and target abundance using qPCR

4. For each sample, set up the following reaction in a 1.5-ml microcentrifuge tube to perform qPCR:

   5 μl reverse transcription product (from step 3)
   2 μl 10 μM mRNA forward primer *or* miRNA forward primer
   2 μl 10 μM mRNA reverse primer *or* universal reverse primer
   23 μl distilled H_2O
   33 μl 2× SYBR Green qRT-PCR mastermix.

   For miRNA samples, use the universal reverse primer shown in Table 2; for target samples, use a standard reverse primer.

   Ensure that the reverse transcription product does not exceed 10% of the total reaction volume; otherwise, it will inhibit the PCR reaction due to carryover of buffer and primers.

5. Aliquot the above mixture into three 0.2-ml PCR tubes at 20 μl per tube. Place on ice until you are ready to run the assay in the RotorGene Q system.

   The three aliquots per sample serve as technical replicates for qPCR.

6. Insert the tubes into the 36-place rotor and affix the locking ring. Place into the machine and run with the following settings:

   Hold 1: 50°C for 20 sec
   Hold 2: 95°C for 600 sec
   Cycling: 95°C for 15 sec, 60°C for 60 sec
   Melt: Ramp from 58°C to 95°C
       Rising by 1°C each step
       Wait for 90 sec of pre-melt conditioning on first step
       Wait for 5 sec for each step afterwards
       Acquire to Melt A on Green
   Uncheck optimize gain before melt on all tubes.

7. Upon completion of program, remove PCR tubes from machine. Download the run file (.rex) and the Rotor-Gene Q Series software; run in virtual mode with serial number 123456 to analyze data on different computers.

Analyze the qPCR results

8. Use the ΔΔCt equation below to determine relative fold change:

$$Relative\ Fold\ Change = 2^{-\Delta\Delta Ct}$$

$$-\Delta\Delta Ct = -\left(\Delta Ct_{Gene\ of\ Interest} - \Delta Ct_{Internal\ Control}\right)$$

$$\Delta Ct = Ct_{Treated} - Ct_{Untreated}$$

Note that this equation assumes equal amplification efficiencies of primers for both the gene of interest and internal control.

9. After calculation of fold change, test for differential expression using the Student's *t*-test on Ct values.

10. Create a scatterplot that compares the fold change by RNA-seq against the fold change by qRT-PCR. Make two separate plots, one for miRNAs and one for targets, with each point on the graph representing a single miRNA or target. Calculate the R^2 correlation value to quantify the strength of the linear relationship between the two methods. An example of such a plot comparing the fold change of eight JA-responsive miRNAs measured by RNA-seq and qRT-PCR is illustrated in Neller et al. (2018).

REAGENTS AND SOLUTIONS

First-strand buffer, 5×

375 mM KCl
15 mM MgCl$_2$
250 mM Tris·Cl, pH 8.3
Store up to 1 year at −20°C

pfu-sso7d buffer, 5×

0.5 mg/ml bovine serum albumin (BSA; Sigma B6917)
50 mM KCl
10 mM MgSO$_4$
50 mM ammonium acetate
150 mM Tris·Cl, pH 10
0.5% (v/v) Triton X-100
Store up to 1 year at −20°C

Sodium citrate buffer, pH 6.4, 0.1 M

Purchase citric acid monohydrate (210.14 g/mol) and sodium citrate dihydrate (294.12 g/mol). Prepare 80 ml of distilled water in a suitable container. Add 0.21 g of citric acid monohydrate to the solution. Add 2.65 g of sodium citrate dihydrate to the solution. Adjust solution to final desired pH using HCl or NaOH. Add distilled water until volume is 0.1 liter.

COMMENTARY

Background Information

Background on miRNA biogenesis and function

miRNAs are non-coding RNAs of ~20 to 24 nt that regulate post-transcriptional gene expression of targets (reviewed in Yu, Jia, & Chen, 2017). As shown in Figure 9, miRNA biogenesis begins with transcription of a *MIR* gene by RNA polymerase II to produce a primary miRNA (pri-miRNA), consisting of a stem-loop region flanked by unstructured arms. In sequential steps, dicer-like 1 RNase (DCL1) excises the stem-loop to form the miRNA precursor (pre-miRNA), then generates a duplex comprising the miRNA and its opposing strand, classically termed microRNA (miRNA). Once in the cytoplasm, the
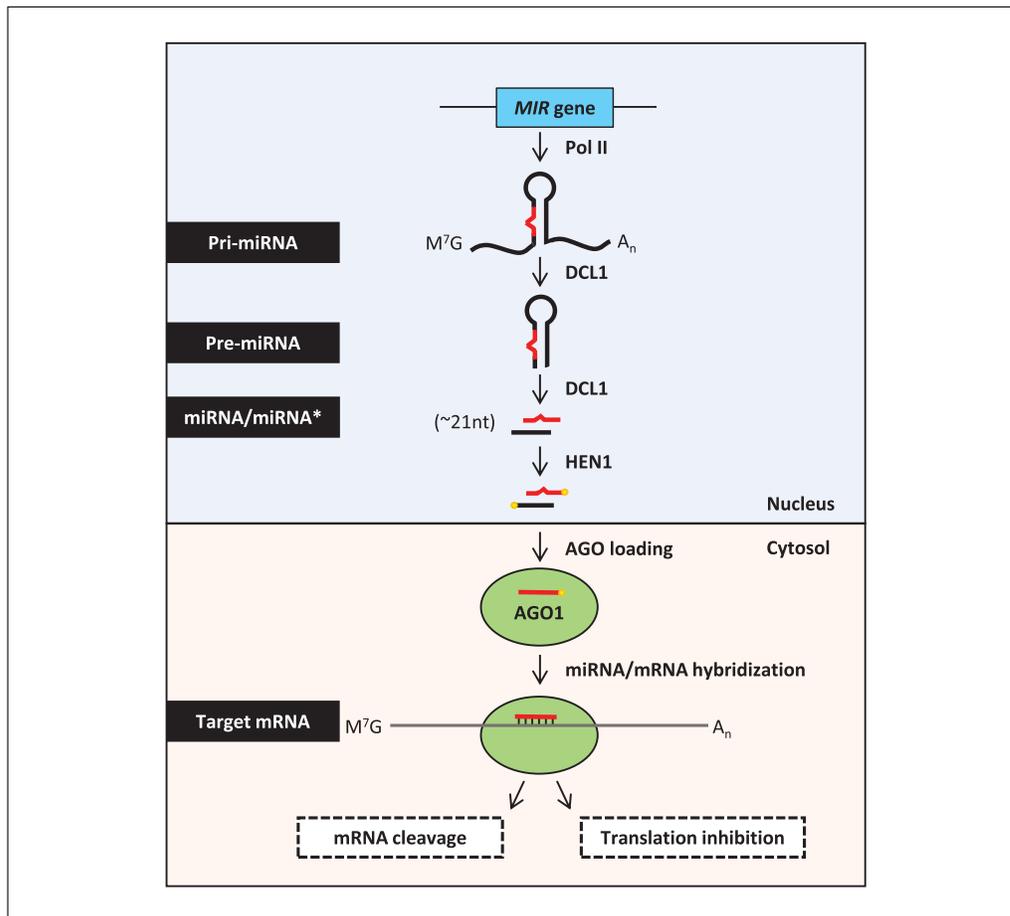
Figure 9 miRNA biogenesis and function. A *MIR* gene is transcribed by Pol II to yield a non-coding, single-stranded transcript that folds back on itself, forming a bulged hairpin flanked by unstructured arms. This primary miRNA (pri-miRNA) is processed mainly by DCL1 to a miRNA precursor (pre-miRNA) and miRNA/miRNA duplex (~21 nt) in sequential steps. The duplex is 3′-end methylated by HEN1 to protect from degradation, then exported to the cytosol. The miRNA guide strand is selected and incorporated into the RNA-induced silencing complex (RISC), which contains an AGO protein, usually AGO1. The RISC binds to a target mRNA on the basis of sequence complementarity and either slices it or inhibits its translation.

RNA-induced silencing complex (RISC) is established upon association of the miRNA with Argonaute 1 (AGO1). Plant miRNAs tend to have high sequence complementarity with targets and act primarily through target mRNA cleavage rather than translational inhibition, the latter being more common in animals. Although the high sequence complementarity of miRNA/target pairs observed in plants is conducive to RNA cleavage, this is not always the case, as miRNA-mediated translational inhibition has been shown in plants for highly complementary miRNA/target pairs.

Intent of the article

The outcome of this article is a candidate set of biologically relevant miRNA/target pairs, along with validations of differential expression and target cleavage. This information may be sought for the purpose of basic research, i.e., to uncover a novel layer of gene expression regulation for the pathway under study. Indeed, plant miRNAs are implicated in numerous abiotic and biotic stress responses, as well as maintenance of normal growth and development (Noman et al., 2017). Alternatively, one may wish to leverage these results in a more applied manner. For example, the use of miRNAs to manipulate transcript abundance is a popular strategy for crop genetic engineering, with several reports demonstrating its successful application in agronomic trait improvement (Djami-Tchatchou, Sanan-Mishra, Ntushelo, & Dubery, 2017).

This article is intended for researchers studying non-model plants. These species tend to lack public genomic resources, most notably a reference genome or transcriptome. Ideally, miRNA prediction is performed at the genomic level, with several prediction

algorithms operating solely on the genome sequence (Rajendiran, Chatterjee, & Pan, 2018). We provide a strategy that considers the unique challenges associated with large-scale bioinformatic analysis of a non-model plant. Specifically, assembly and annotation of a reference transcriptome is easier, faster, and less expensive than that of a genome. The former also requires less specialized knowledge, due in large part to the availability of integrated, user-friendly tools aimed at biologists with limited bioinformatic experience. It is important to note that the use of genome or transcriptome references from related plant species is not recommended. This is because plant miRNAs tend to be highly species-specific, and unlike animals, miRNA precursors are generally not conserved across plants (Bartel, 2004). For plant miRNAs that are conserved, this usually occurs at the level of the mature miRNA (Bartel, 2004). Even so, we advise against miRNA prediction based solely on sequence comparison against mature miRNAs of related species, as the identification of a suitable miRNA precursor is an important feature used to reduce false positive predictions. Therefore, the approach of this article is to create and use resources that are specific to the plant under study.

Strengths and limitations of the procedure

Due to the tendency of plant miRNAs to be species-specific, an advantage of our procedure is that miRNAs are predicted using a transcriptomic reference generated for the species under study. Additionally, the integration of both miRNA and target differential expression in the procedure described in this article provides a filter to identify the most biologically relevant interactions for downstream analysis. This approach assumes that miRNAs and targets whose expression levels change in response to the treatment variable are more likely to be important mediators of the response than those pairs whose expression levels do not change significantly. This is a reasonable assumption given that the expression of an miRNA and its mRNA target tend to be correlated, an attribute that has been exploited successfully by us and others to discover relevant miRNA/target pairs in plants (Ji et al., 2018; Ma et al., 2018; Neller et al., 2018; Ye, Wang, & Wang, 2016). Therefore, an investigation of differentially expressed miRNAs without considering target expression only reveals half of the story. Our use of paired small RNA and mRNA samples in this procedure enables downstream

application of correlation methods to further refine miRNA/target relationships.

Restriction of the analysis to targets that are differentially expressed has a limitation in that it may filter out miRNAs acting through translational inhibition rather than transcript cleavage. However, since this mechanism of action is not observed frequently in plants, the limitation is not critical unless the reader is specifically interested in miRNAs that operate in this manner. Furthermore, the two modes of miRNA action may be difficult to distinguish in plants, as miRNA-induced cleavage occurs on targets undergoing active translation (reviewed in Yu et al., 2017). Another limitation of our procedure is that validation by the modified RLM-RACE used in this article only provides information on the presence of cleaved targets, not their relative abundance, and it may be unsuccessful in the case of low-abundance miRNAs or targets, where less cleavage product is available for detection. A high-throughput equivalent to this procedure is degradome sequencing (Lin, Chen, & Lu, 2019). By incorporating RNA-seq, this method enables detection of all cleaved targets in a sample and their relative abundance. It is less economical and requires extensive data analysis, but the combination of small RNA-seq and degradome-seq is highly informative; see Ji et al. (2018) for a recent implementation of this strategy.

Comparison of the bioinformatic workflow with current methods

The workflow presented in this article is based on our experience with available software options. We have prioritized qualities of open access, user friendliness, in-depth documentation, and smooth integration. Some components of our workflow are more advanced than others. For the tasks of de novo transcriptome assembly and annotation, we highly recommend Trinity and its companion software Trinotate and TransDecoder. Additionally, we suggest the use of scripts packaged with Trinity to facilitate differential expression and GO enrichment analyses. Each of these tasks requires an advanced level of expertise that can overwhelm a novice user, resulting in incorrect application of methods. For this reason, we view the integration and guidance provided by Trinity developers as highly advantageous. However, other options do exist for de novo transcriptome assembly. For a plant-focused summary of these tools and other resources, see Geniza and Jaiswal (2017). Additionally, refer to Honaas et al. (2016) for a comparative

analysis of transcriptomes generated from different assemblers for the model plants rice and *Arabidopsis*. It is also possible to bypass de novo transcriptome assembly completely by performing Iso-seq, a relatively new implementation of long-read technology that has been used in plants (An, Cao, Li, Humbeck, & Wang, 2018). Regardless of the chosen strategy, note that Trinotate and TransDecoder can be used for annotating any transcriptome as long as the required inputs are provided.

Other aspects of the workflow are more amenable to user customization. For example, there are numerous options available for miRNA prediction and target identification (reviewed in Rajendiran et al., 2018). Recently, supervised machine learning was used to predict miRNAs in a reference-free manner (Vitsios et al., 2017). Although promising, this approach was more successful in predicting miRNAs for animals than plants. Furthermore, it is unlikely to outperform reference-based prediction for a non-model plant, as the user must provide training data consisting of known miRNAs or instead use the 'universal plant' model. Other options for user customization of our workflow are at the level of transcript quantification and differential expression analysis. We used the alignment-based method RSEM for transcript quantification, but alignment-free methods such as Kallisto and Salmon are also popular due to their improved speed. Although alignment-free and alignment-based methods are comparable in accuracy for standard investigations like protein-coding mRNA quantification, note that alignment-based methods perform better when quantifying small or low-abundance RNAs (Wu, Yao, Ho, Lambowitz, & Wilke, 2018). The reader may also wish to investigate alternatives for differential expression analysis, with popular options including DESeq2 and Voom/Limma. The Trinity accessory scripts used in our workflow support these various programs/packages for transcript quantification and differential expression analysis, thereby accommodating differences in experimental design and user preference. For a comparison of RNA-seq mapping methods (both alignment-free and alignment-based) and differential expression tools, see Costa-Silva, Domingues, and Lopes (2017).

Extended bioinformatic analysis

With the paired small RNA and mRNA samples as used in our workflow, the investigator is equipped to perform advanced correlation analysis for obtaining greater insight into miRNA/target interactions. In our study, we computed the Pearson correlation coefficient (PCC) for the expression of each miRNA and its target(s) and imposed a PCC cut-off to filter the set of candidate pairs. The PCC ranges from -1 to $+1$, indicating perfect negative and positive linear association, respectively. There is a tendency in literature to retain only negatively correlated miRNA/target pairs. This derives from the rationale that a cleavage-inducing miRNA acting on its target in absence of other influences leads to reduced target expression due to RNA degradation. However, we and others have observed and validated positively correlated miRNA/target pairs. This dynamic can arise from miRNA-mediated spatial restriction of the target (Kawashima et al., 2009; Kidner & Martienssen, 2004; Levine, McHale, & Levine, 2007; Nikovics et al., 2006). It may also indicate that the miRNA functions in a 'buffering' capacity, minimizing changes in target expression caused by other interacting factors (Wu, Shen, & Tang, 2009). For these reasons, we do not recommend restricting analysis to negatively correlated miRNA/target pairs. If readers are interested in generating an advanced miRNA/target interaction network, we direct them to reviews on the various mathematical models and integrated approaches used (Carroll, Goodall, & Liu, 2014; Muniategui, Pey, Planes, & Rubio, 2013). For simple visualization of predicted miRNA/target interactions, we recommend import of results from this workflow into Cytoscape (Shannon et al., 2003).

Wet-lab procedures

RNA extraction

High-quality RNA is essential for both RNA-seq and the validations used in this article. It consists of RNA that is primarily free of degradation by cellular nucleases and lacks contamination by genomic DNA. Extraction of RNA begins by grinding the tissue sample and solubilizing its contents. Solubilization buffers containing guanidinium compounds protect against nucleases and aid in breakdown of the cell membrane (Chomczynski, 1993). Following solubilization, the user can continue with a reagent-based extraction, such as with RNAzol, or switch to purification with silica columns. RNAzol is advantageous as it enables convenient isolation of the small RNA fraction and contains additives to reduce genomic DNA contamination. Alternatively, silica-based purification allows much faster RNA isolation

and on-column DNase treatment to remove genomic DNA; however, a major drawback of this technology is its often-poor yield.

RLM-RACE

5′ RACE was originally developed to map the +1 transcription start site of mRNAs (Sambrook & Russell, 2006). In this application, the mRNA was reverse transcribed with an oligo d(T) primer, then terminal deoxytransferase (TdT) was used to add multiple nucleotides to the 3′ end of the cDNA (known as 'tailing') to produce an adapter sequence. RLM-RACE forgoes tailing and instead ligates a 5′ RNA adapter directly to an mRNA pool that has been phosphatase-treated and decapped to select for full-length mRNAs. This is ideal for the original application of the method but not for validating miRNA-induced cleavage events, since the targeted mRNA is sliced. To modify this procedure for detecting cleaved products, we omitted the selection of full-length mRNAs, resulting in any exposed 5′ phosphate in the mRNA pool becoming ligated to the RNA adapter. Subsequent amplification with PCR and cloning with Gibson assembly allows the identification of 5′ cut sites of specific mRNA targets.

RLM-RACE is superior to 5′ RACE for detecting miRNA-induced cleavage. T4 RNA ligase is efficient at adding an RNA adapter to the 5′ end of mRNA, while TdT used to tail the cDNA in 5′ RACE can add nucleotides to ss-DNA, dsDNA, and, at a lower efficiency, RNA, meaning there is less specificity for the cDNA of cleaved mRNA. Additionally, RLM-RACE avoids potential artifacts caused by the reverse transcriptase stalling during cDNA synthesis. As described above, degradome-seq is a high-throughput equivalent to the modified RLM-RACE used in this article, enabling relative quantification of all cleaved targets. To quantify miRNA-directed repression regardless of whether it derives from transcript cleavage or translational inhibition, a transient dual luciferase assay has been optimized for use in Nicotiana benthamiana (Moyle et al., 2017).

qRT-PCR and stem-loop qRT-PCR

qRT-PCR is a common and rapid method for relative RNA quantification. The transcript of interest is quantified in both the treated and untreated sample, with the change in expression normalized to that of an internal control transcript in each sample to account for differences in amount of starting RNA. Internal controls are often referred to as 'housekeeping' genes for their stable expression across various treatments. In this article, we use qRT-PCR and stem-loop qRT-PCR to validate expression of target mRNAs and miRNAs, respectively. Design of primers for standard qRT-PCR is relatively straightforward: the two primers should produce a product of 150 to 200 bp. To identify and reduce the impact of genomic DNA contamination in the RNA sample, primers should span an intronic region if one is known to exist. Stem-loop qRT-PCR, developed by Chen et al. (2005), uses a stem-loop primer complementary to the last six bases on the 3′ end of the miRNA. This increases the length and melting point of the PCR product to make it compatible with standard cycling. Therefore, specificity of the PCR reaction is conferred mainly by the forward primer, which spans most of the miRNA sequence.

We use SYBR Green to measure fluorescence during qPCR, but specificity can be increased by substituting sequence-specific hydrolysis probes such as TaqMan (Applied Biosystems) or Universal ProbeLibrary (UPL, Roche Diagnostics). The probes anneal to single-stranded DNA and emit fluorescence only upon DNA polymerase-induced cleavage, which results from 5′ to 3′ exonuclease activity of the polymerase as it extends the primer. Use of these probes reduces background fluorescence due to primer dimers and increases specificity, as the probe binds between primer annealing sites. UPL probes provide increased specificity by incorporating locked nucleic acids, which are modified nucleotides with ribose rings stabilized in an ideal conformation for Watson-Crick base pairing. For a protocol utilizing UPL stem-loop qRT-PCR to quantify low-abundance plant miRNAs, see Varkonyi-Gasic, Wu, Wood, Walton, and Hellens (2007). Although beneficial, the high cost of hydrolysis probes likely excludes them from use in initial screening validations, but their incorporation may be worthwhile for characterization of a few key miRNAs.

Critical Parameters

High-quality RNA is an essential input for both RNA-seq and the validations performed in this article. General plant health is an important contributor to RNA quality, with lack of light or nutrients resulting in lesser-quality RNA. If the treatment under study is intended to elicit a stress response, the strength and duration of treatment must be optimized to avoid impacting overall RNA quality. When harvesting plants, tissue should be flash-frozen in

Table 3 Troubleshooting RNA Extraction

Problem	Possible cause	Solution	Notes
RNA degradation (28S intensity <18S) RNA integrity number (RIN) <8	Samples thawing during extraction Contamination of samples Unhealthy plants	Keep samples at liquid nitrogen temperatures during grinding Homogenize with RNAzol quickly Do not talk over tubes Use filter tips and clean glassware Resuspend pellets in warm RNA storage buffer Keep plants watered and fed	RNA extraction requires a level of manual dexterity and practice to achieve optimal results. Degradation is evident if the 28S rRNA band appears fainter than the 18S on an agarose gel or if the RIN is <8. At the initial stage, degradation can occur if the ground leaf powder is not kept at liquid nitrogen temperatures and if the powder is not solubilized in the RNAzol fast enough.
Genomic DNA contamination	Contaminants introduced from pellet in Support Protocol, step 3 High DNA or polysaccharide content in plant tissue	Take less supernatant for RNA precipitation Reduce amount of tissue in RNA extraction Use the optional step of adding 4-bromoanisole (see RNAzol manual) DNase-treat RNA	Genomic DNA contamination can be detected as high-molecular-weight bands or smears when total RNA is separated on an agarose gel. As well, qRT-PCR reactions can be separated on an agarose gel after the reaction has finished; products of unexpected size indicate genomic contamination if qRT-PCR primers were designed to span intron/exon junctions.

liquid nitrogen and processed immediately. Use of pre-chilled tools and tubes prevents frozen tissue from melting, thereby limiting nuclease activity. If liquid nitrogen is unavailable, tissue can be preserved in saturated ammonium sulfate solution, such as RNAlater (Ambion). Note that the lysate is stable upon solubilization in RNAzol and can be stored long-term at −20°C. Once extracted, RNA is susceptible to degradation, by nucleases and resulting from hydrolysis under basic conditions. Both factors can be controlled by resuspending RNA pellets in RNA storage buffer, which contains DTT to inactivate ribonucleases and citrate buffer to reduce pH and chelate metal ions.

Sufficient computational power and memory are required to perform RNA-seq analysis. Raw data and output files must be stored, and the workflow must be able to complete in a reasonable timeframe. We performed all analysis for the full-scale job using a personal server with 64 Gb RAM and a 4 Tb hard drive. If suitable infrastructure is not available, public options for researchers include Galaxy and CyVerse (originally iPlant Collaborative; Merchant et al., 2016) for both high-performance computing and data storage. We introduce the user to Galaxy in this article.

Certain aspects of the RNA-seq experimental design and bioinformatic workflow are essential to include. If differential expression analysis is performed, both biological replicates and strand-specific reads must be incorporated. The former allow for an assessment of sample variability, while the latter ensures accurate transcript quantification. Note that many peer-reviewed journals now require a minimum of three biological replicates for inclusion of RNA-seq differential expression analysis. It is also essential to quality control the raw reads prior to analysis. Contaminating adapter sequences create artifacts in both the assembled transcriptome and

Neller et al.

Table 4 Troubleshooting RLM-RACE

Problem	Possible cause	Solution	Notes
No bands on agarose gel after nested PCR	Poor RNA quality Low abundance mRNA Degraded RNA adapter RNA still bound to cDNA Primer design	Check RNA on agarose gel or bioanalyzer Increase amount of starting RNA by 50%-100% Use a fresh aliquot of RNA adapter Treat cDNA with RNase H after reverse transcription Check primer design Use positive control (miR156/SPL mRNA)	It is imperative to have a working positive control for RLM-RACE. For plants, miR156 and its target, SQUAMOSA-PROMOTER BINDING PROTEIN-LIKE (SPL) are both abundant and easy to detect. If the positive control gives the product of expected size on a gel, then errors in adapter ligation and losses in mRNA purification can be ruled out.
Unexpected and random cleavage sites in target mRNA	Poor RNA quality Low-abundance miRNA miRNA is non-functional	Check RNA on agarose gel or bioanalyzer Increase amount of starting RNA by 50%-100% Scrutinize bioinformatics predictions Choose another miRNA to validate	Cleavage at unexpected sites suggests that the miRNA is not cleaving the expected target, or that this small RNA is not a functional miRNA (i.e., false positive). In some cases, we observed randomly distributed cleavage surrounding the expected site. This suggested that the limit of detection had been reached; background levels of mRNA degradation were as likely as miRNA cleavage.

small RNA sequences, and low-quality bases introduce sequence errors. Both issues interfere with transcriptome assembly, differential expression analysis, and miRNA prediction.

Factors affecting success of wet-lab validations are also important to consider. Low-abundance miRNAs (<10 CPM) and targets (<100 TPM) are difficult to detect. Cleaved mRNA is quickly degraded in the cell, so a low initial abundance further inhibits detection by RLM-RACE. Similarly, low abundance results in undetectable or variable Ct values during qRT-PCR. Another critical factor for a successful qRT-PCR experiment is the selection of appropriate internal controls. A preliminary test should be performed to ensure that expression is stable for the treatment under study. Two popular programs that aid in the selection of internal controls are NormFinder (Aanes et al., 2014) and geNorm (Vandesompele et al., 2002). These programs use pairwise comparisons of expression data to rank the stability of reference genes. Both are available as Microsoft Excel plugins, which allows analysis without extensive bioinformatics knowledge. However, these programs are unable to process large transcriptome datasets with ease. We reduced the list of input reference genes by calculating the mean, standard deviation, and relative standard deviation of abundance for each transcript across all samples using Excel. Relative standard deviation is calculated by multiplying standard deviation by 100 and dividing by the mean. Transcripts were sorted first by lowest relative standard deviation and then by highest mean expression level. Such transcripts would vary little in level across different treatments and would be of sufficient abundance to detect reliably. The reduced gene list was used as input for NormFinder. The final list of candidate reference genes was verified with qRT-PCR across the different treatments tested.

Troubleshooting

For problems concerning any portion of the bioinformatic workflow, we refer the reader to the appropriate program guides at the links

Table 5 Troubleshooting qRT-PCR

Problem	Possible cause	Solution	Notes
No amplification	Critical reagent missing/degraded PCR cycling/primer design	Use fresh reagents Check PCR program, primers Use positive control	No amplification is often a case of technical error, either from missing reagents or incorrectly used equipment
Variable/ unpredictable amplification	Poor RNA quality Low abundance miRNA/mRNA False positive miRNA/mRNA pair	Increase amount of starting RNA	Increasing input of RNA by 50%-100% can help detect low-abundance miRNA/mRNA pairs; however, this may cause the internal control to amplify earlier as well
Very early amplification	Genomic DNA contamination Too much input RNA	Check qRT-PCR products on agarose gel for larger than expected products Reduce amount of starting RNA	With appropriate primer design–spanning introns, genomic DNA–based product will be larger than true reverse-transcribed RNA. Too much RNA can skew Ct values; both internal control and experimental samples should amplify between cycles 15 and 20

provided. Ensure that each program is installed correctly, including any required dependencies, and that it is accessible from the user's `$PATH`. Additionally, ensure that the full file path (`/path/to/file.ext`) is provided for input on the command line; otherwise, the file may not be located by the program. For troubleshooting common issues that may be experienced during wet-lab procedures, refer to Tables 3 to 5.

Anticipated Results

Bioinformatic analysis

For an overview of expected results, we refer the reader to the accompanying Supporting Information, which provides all output files referenced in this article. Beginning with raw RNA-seq data, the user will perform read cleaning, de novo transcriptome assembly and annotation, miRNA prediction and annotation, differential expression analysis, miRNA/ target prediction, and functional enrichment analysis.

Wet-lab procedures

When visualizing extracted RNA on a gel, the following should be observed. For total RNA, a distinctive pattern forms due to the high abundance of rRNA: 28S produces the brightest band, followed by 18S and 5S. Equal intensity of 28S and 18S is acceptable, while a brighter 18S indicates degradation. Enriched mRNA should appear as banding or smearing above the 1 Kb marker of an RNA-specific ladder, while the small RNA-enriched fraction produces a smear below the 200 bp marker. In RLM-RACE, nonspecific amplification of the cDNA pool should produce an indistinct smear on an agarose gel. The first round of gene-specific amplification may produce multiple bands, a smear, or no visible product. Following nested PCR, the expected band should be observed, either alone or easily distinguished from a small number of nonspecific bands. For qRT-PCR, relative fold changes between treated and untreated samples may not coincide 1:1 with the RNA-seq results. However, there should be a strong linear correlation between the two methods ($R^2 > 0.8$).

Time Considerations

Bioinformatic analysis

Using the sample data and reference files provided in this article, each protocol can be

Neller et al.

completed comfortably within an hour. Tasks requiring extensive computational resources and run time are transcriptome de novo assembly, BLAST homology searches, and any read-mapping procedures. Completion times will vary depending on the user's resources. Completion times for the full-scale job on our relatively modest server (64 Gb RAM, 16 CPUs) ranged from several hours (read-mapping) to several days (de novo assembly and BLAST).

Wet-lab procedures

We recommend implementing a consistent planting schedule so that minimal time is spent waiting for plants to grow. Once plants are available and the experimental treatment applied, samples should be flash-frozen in bulk and processed in smaller batches. The RNA extraction protocol requires 1 to 2 days for a single batch of 8 to 16 samples. Validation of 5 to 10 targets with RLM-RACE will require at least 4 weeks. The limiting step is PCR optimization, which may require testing of multiple primer pairs. Colony screening should take 1 to 2 days, with Sanger sequencing results likely available in 2 to 3 days. For qRT-PCR, reverse transcription can be performed comfortably on 12 to 24 samples within 1 day. Assuming use of a 36-well rotor, 36 samples can be assessed by qPCR in one batch. As an example, this could be split between six 'gene targets' (e.g., 5 miRNAs + 1 internal control miRNA) × 3 technical replicates × 2 conditions. Sample setup and qPCR can be performed within 1 day for a single batch. Note that optimization of starting RNA amount and primer design will likely be required for low-abundance miRNA/target pairs.

Acknowledgements

Literature Cited

Aanes, H., Winata, C., Moen, L. F., Østrup, O., Mathavan, S., Collas, P., ... Aleström, P. (2014). Normalization of RNA-sequencing data from samples with varying mRNA levels. *PloS One*, *9*, 1–7. doi: 10.1371/journal.pone.0089158.

An, D., Cao, H., Li, C., Humbeck, K., & Wang, W. (2018). Isoform sequencing and state-of-art applications for unravelling complexity of plant transcriptomes. *Genes (Basel)*, *9*, 43. doi: 10.3390/genes9010043.

Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, *116*, 281–297. doi: 10.1016/S0092-8674(04)00045-5.

Carroll, A. P., Goodall, G. J., & Liu, B. (2014). Understanding principles of miRNA target recognition and function through integrated biological and bioinformatics approaches. *Wiley Interdisciplinary Reviews, RNA*, *5*, 361–379. doi: 10.1002/wrna.1217.

Chen, C., Ridzon, D. A., Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., ... Guegler, K. J. (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Research*, *33*, 1–9. doi: 10.1093/nar/gni178.

Chomczynski, P. (1993). A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples. *Biotechniques*, *15*, 532–4, 536–537.

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS One*, *12*, e0190152. doi: 10.1371/journal.pone.0190152.

Djami-Tchatchou, A. T., Sanan-Mishra, N., Ntushelo, K., & Dubery, I. A. (2017). Functional roles of microRNAs in agronomically important plants—Potential as targets for crop improvement and protection. *Frontiers in Plant Science*, *8*, 378. doi: 10.3389/fpls.2017.00378.

Geniza, M., & Jaiswal, P. (2017). Tools for building de novo transcriptome assembly. *Current Plant Biology*, *11–12*, 41–45. doi: 10.1016/j.cpb.2017.12.004.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., ... Nekrutenko, A. (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, *15*, 1451–1455. doi: 10.1101/gr.4086505.

Honaas, L. A., Wafula, E. K., Wickett, N. J., Der, J. P., Zhang, Y., Edger, P. P., ... dePamphilis, C. W. (2016). Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome. *PloS One*, *11*, e0146062. doi: 10.1371/journal.pone.0146062.

Ji, Y., Chen, P., Chen, J., Pennerman, K., Liang, X., Yan, H., ... Huang, L. (2018). Combinations of small RNA, RNA, and degradome sequencing uncovers the expression pattern of microRNA–mRNA pairs adapting to drought stress in leaf and root of *Dactylis glomerata* L. *International Journal of Molecular Sciences*, *19*, 3114. doi: 10.3390/ijms19103114.

Kawashima, C. G., Yoshimoto, N., Maruyama-Nakashita, A., Tsuchiya, Y. N., Saito, K., Takahashi, H., & Dalmay, T. (2009). Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different

cell types. *Plant Journal*, *57*, 313–321. doi: 10.1111/j.1365-313X.2008.03690.x.

Kidner, C. A., & Martienssen, R. A. (2004). Spatially restricted microRNA directs leaf polarity through ARGONAUTE1. *Nature*, *428*, 81–84. doi: 10.1038/nature02366.

Levine, E., McHale, P., & Levine, H. (2007). Small regulatory RNAs may sharpen spatial expression patterns. *PloS Computational Biology*, *3*, 2356–2365. doi: 10.1371/journal.pcbi.0030233.

Lin, S. S., Chen, Y., & Lu, M. J. (2019). Degradome sequencing in plants. *Methods in Molecular Biology*, *1932*, 197–213. doi: 10.1007/978-1-4939-9042-9_15.

Ma, C., Yang, J., Cheng, Q., Mao, A., Zhang, J., Wang, S., ... Wen, C. (2018). Comparative analysis of miRNA and mRNA abundance in determinate cucumber by high-throughput sequencing. *PloS One*, *13*, e0190691. doi: 10.1371/journal.pone.0190691.

Mathioni, S. M., Kakrana, A., & Meyers, B. C. (2017). Characterization of plant small RNAs by next generation sequencing. *Current Protocols in Plant Biology*, *2*, 39–63. doi: 10.1002/cppb.20043.

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., & Antin, P. (2016). The iPlant collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences. *PloS Biology*, *14*, e1002342. doi: 10.1371/journal.pbio.1002342.

Moyle, R. L., Carvalhais, L. C., Pretorius, L.-S., Nowak, E., Subramaniam, G., Dalton-Morgan, J., ... Schenk, P. M. (2017). An optimized transient dual luciferase assay for quantifying MicroRNA directed repression of targeted sequences. *Frontiers in Plant Science*, *8*, 1631. doi: 10.3389/fpls.2017.01631.

Muniategui, A., Pey, J., Planes, F. J., & Rubio, A. (2013). Joint analysis of miRNA and mRNA expression data. *Briefings in Bioinformatics*, *14*, 263–278. doi: 10.1093/bib/bbs028.

Neller, K. C. M., Klenov, A., Guzman, J. C., & Hudak, K. A. (2018). Integration of the pokeweed miRNA and mRNA transcriptomes reveals targeting of jasmonic acid-responsive genes. *Frontiers in Plant Science*, *9*, 589. doi: 10.3389/fpls.2018.00589.

Neller, K. C. M., Klenov, A., & Hudak, K. A. (2016). The pokeweed leaf mRNA transcriptome and its regulation by jasmonic acid. *Frontiers in Plant Science*, *7*, 1–13. doi: 10.3389/fpls.2016.00283.

Nikovics, K., Blein, T., Peaucelle, A., Ishida, T., Morin, H., Aida, M., & Laufs, P. (2006). The balance between the MIR164A and CUC2 genes controls leaf margin serration in Arabidopsis. *Plant Cell*, *18*, 2929–2945. doi: 10.1105/tpc.106.045617.

Noman, A., Fahad, S., Aqeel, M., Ali, U., Amanullah Anwar, S., ... Zainab, M. (2017).

miRNAs: Major modulators for crop growth and development under abiotic stresses. *Biotechnology Letters*, *39*, 685–700. doi: 10.1007/s10529-017-2302-9.

Rajendiran, A., Chatterjee, A., & Pan, A. (2018). Computational approaches and related tools to identify microRNAs in a species: A bird's eye view. *Interdisciplinary Sciences: Computational Life Sciences*, *10*, 616–635. doi: 10.1007/s12539-017-0223-x.

Sambrook, J., & Russell, D. W. (2006). Rapid amplification of 5′ cDNA Ends (5′-RACE). *CSH Protocols*, *1*, pii: Pdb.prot3989. doi: 10.1101/pdb.prot3989.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*, 2498–2504. doi: 10.1101/gr.1239303.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., & Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, *3*, RESEARCH0034. doi: 10.1186/gb-2002-3-7-research0034.

Varkonyi-Gasic, E., Wu, R., Wood, M., Walton, E. F., & Hellens, R. P. (2007). Protocol: A highly sensitive RT-PCR method for detection and quantification of microRNAs. *Plant Methods*, *3*, 12. doi: 10.1186/1746-4811-3-12.

Vitsios, D. M., Kentepozidou, E., Quintais, L., Benito-Gutiérrez, E., Van Dongen, S., Davis, M. P., ... Enright, A. J. (2017). Mirnovo: Genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Research*, *45*, e177. doi: 10.1093/nar/gkx836.

Voytas, D. (2000). Agarose gel electrophoresis. *Current Protocols in Molecular Biology*, *51*, 2.5A.1–2.5A.9.

Wu, C. I., Shen, Y., & Tang, T. (2009). Evolution under canalization and the dual roles of microRNAs—A hypothesis. *Genome Research*, *19*, 734–743. doi: 10.1101/gr.084640.108.

Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M., & Wilke, C. O. (2018). Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, *19*, 510. doi: 10.1186/s12864-018-4869-5.

Ye, B., Wang, R., & Wang, J. (2016). Correlation analysis of the mRNA and miRNA expression profiles in the nascent synthetic allotetraploid Raphanobrassica. *Scientific Reports*, *6*, 37416. doi: 10.1038/srep37416.

Yu, Y., Jia, T., & Chen, X. (2017). The "how" and "where" of plant microRNAs. *The New Phytologist*, *216*, 1002–1017. doi: 10.1111/nph.14834.