Small RNAs in Duckweeds

16

Paul Fourounjian

Abstract

Within eukaryotic genomes, there are several types of small RNAs including sn, sno, si, and miRNAs. With respect to the Lemnaceae family, the vast majority of the research to date has been conducted in *Spirodela polyrhiza*, focused mainly on the miRNAs. This research consists of three small RNA-sequencing experiments in strains from China, Germany, and the USA, with each experiment identifying conserved miRNAs and predicting novel miRNAs and targets. While the novel miRNAs and recently discovered miRNAs fluctuated, the family size and expression of well-known miRNA families was consistent between the three experiments. While miRBase likely contains many incorrect annotations, these miR-NAs were annotated according to strict criteria and analyzed for the miRBase high confidence list. They were further characterized through degradome sequencing, which confirmed half of the conserved miRNAs and a third of the novel. Finally, *Spirodela polyrhiza* has a surprisingly low abundance of 24nt sRNAs, which are required to suppress transposon proliferation.

As scientists moved from sequencing the ϕX174 virus in 1977 to prokaryote genomes, simple eukaryotes, and then the first plant, *Arabidopsis thaliana* in the year 2000, they saw that these larger complex genomes were made of so much more than genes. We now know that eukaryotic genomes contain a host of structural repeats such as the centromere and telomere regions. There are also large stretches of tandem repeats, also called satellite DNA. Then, there are the virus-like transposable elements that are often copied and spread across the genome. Many of the transcribed RNA sequences are small RNAs like small interfering, micro, and small nucleolar RNAs (si, mi, and snoRNAs) that bind to protein complexes to regulate gene expression and assemble ribosomes. Larger RNA transcripts include long non-coding RNAs and the high copy number ribosomal and transfer RNAs (lnc, r, and tRNAs) that translate mRNAs to proteins. Each genome also contains plenty of pseudogenes, which are non-functional due to mutations. Finally, the genome contains the protein-coding genes themselves, with all their introns, exons, cis- and trans-regulatory elements and terminators, which are 2% of the human genome and roughly 20% of a typical angiosperm genome, with wide variation due to genome size differences.

Within this genome, there are several types of transcribed RNAs, with the longer varieties including m, r, t, and lncRNA. While the first three types are well characterized, long

P. Fourounjian (✉)
Waksman Institute of Microbiology, Rutgers University, Piscataway 08854, USA
e-mail: pjf99@scarletmail.rutgers.edu

non-coding RNA (lncRNA) wasn't discovered until 1990 (Brannan et al. 1990). These spliced and polyadenylated RNAs function in epigenetic regulation, the generation and sequestration of miRNAs, and various other functions. While most of the studies have been run in animals, thousands of lncRNAs have been annotated in plant genomes, including IPS1, which sequesters miR399 with a non-cleavable target bulge in response to phosphate starvation across many plant species (Franco-Zorrilla et al. 2007).

The small RNAs in plants include sno, si, and miRNAs, with the snoRNAs evolutionarily conserved back to Archaea. They are produced from their own RNA precursors, or introns, which are cleaved by endonucleases and trimmed by exonucleases, until only the protein bound 60–250 bp snoRNA remains; they then guide the protein complex's methylation and pseudouridylation of rRNAs in the nucleolus. It is even hypothesized that snoRNAs gave rise to miRNAs based on their similarity in processing including some overlap of enzymes, their similar hairpin structure, and combination of function (Scott and Ono 2011). There have been reports of snoRNAs with miRNA-like characteristics, and vice versa, and even small RNAs with complete sno and miRNA function in animals, plants, and yeast. In plants, both miRNAs and siRNAs are cut to 22 and 21nt lengths by dicer-like proteins 1 and 4, respectively, and loaded onto Ago1 in the RISC, with the main difference being that an RNA hairpin is processed into a miRNA for mRNA gene suppression, while a dsRNA is diced into many siRNAs for pathogen gene silencing.

When the *Spirodela polyrhiza* genome was published in 2014, prediction programs were able to detect miRNA precursors through homologous sequences and RNA folding software (Wang et al. 2014a). In strain 7498, all miRBase plant mature sequences were mapped back to the genome, and flanking sequences analyzed by RNAfold and miRCheck (Denman 1993; Jones-Rhoades and Bartel 2004). The search predicted 413 miRNAs belonging to 93 families. This survey based on DNA sequencing aimed to provide all possible miRNA genes, for comparison to other plant genomes, with the eventual aim of detecting their activity in later RNA-seq experiments.

The earliest attempt at sequencing and analyzing *S. polyrhiza* miRNAs predated the published genome. This experiment, run at Peking University Shenzhen Graduate School, was run on strain LT5a, isolated from Lake Tai, using three populations grown in SH media for 1, 3, and 5 days under control conditions. Using 18-31nt sRNA on a HiSeq 2000 Illumina platform, they sequenced 24 million reads, 3.5 of which matched conserved miRNAs in miRBase, and 7.6 million that were not annotated in GenBank or Rfam. These 7.6 million reads were analyzed by the MIREAP program and validated by Mfold to identify 41 predicted novel miRNAs (Zuker 2003). A summary of this and the other small RNA-seq experiments is available in Table 16.1.

In strain 9509, conserved and novel miRNAs were identified through small RNA-sequencing and an analysis of read count and distribution (Michael et al. 2017). The study used 10 sRNA libraries from a SOLiD5500 sequencer, aligned to the genome allowing 1 mismatch, and then annotated if the candidate has a stable hairpin structure, sufficient miR reads, more than 1 miR* read, and a 2 or 3 nt 3′ overhang (Table 16.1). They identified conserved miRNAs by checking for a strong BLAST homology to not only the mature, but also hairpin structures in miRBase. Next they used the program TargetFinder with a cutoff score of 4 to identify the predicted targets (Fahlgren and Carrington 2010). These transcription and structural requirements lead to the prediction of 59 conserved miRNAs in 22 families, and 29 novel miRNAs, with 29 of the conserved and 25 of the novel miRNAs being predicted to regulate 991 mRNA targets.

Alongside the miRNA prediction, they were able to predict trans-acting siRNAs (tasiRNAs), from the sRNA library using previously established criteria (Howell et al. 2007; Johnson et al. 2009). Reads matching cDNA and the corresponding genomic regions had miRNA results filtered out, and then, 50nt candidate transcripts were required to have over 100 reads, with over

Table 16.1 Summary of sRNA-sequencing experiments

Strain	LT5a	7498	9509
Conditions	Control (SH media, 16 h days, 23 °C)	Control, heat, cold, abscisic acid, copper, kinetin, nitrate, sucrose	Control, abscisic acid
# reads	25 million	32 million	N/A
# conserved miRNAs	158	58	59
# novel miRNAs	41	14	29
# targets	N/A	162	991
# DE miRNAs	N/A	15	12

70% being 21nt in length. These are sufficient to distinguish randomly degraded transcripts from mRNAs that had been transcribed into dsRNA and then diced into 21nt tasiRNAs. TargetFinder was then used with a cutoff of 6 and a requirement of two miRNA bind sites to identify the targeted genes. This search yielded two cleaved TAS3 genes, and the miR393 targeting another putative TAS gene that was also found in oil palm and banana.

The most recent miRNA survey started with strain 7498 grown in three replicate flasks of eight growth conditions: control, cold, heat, abscisic acid, copper, kinetin, nitrate, and sucrose stimuli. After harvest, RNA extraction, and size selection, 32 million reads of the 24 libraries were sequenced on the SOLiD5500 platform and mapped to the genome (Table 16.1). These results were filtered against *Brachypodium distachyon* non-coding RNAs, with miRNAs removed, and analyzed in miRPlant (An et al. 2014). Criteria required a miRPlant score greater than 3.0, over 20 miR reads and at least 1 miR* read. This yielded 58 conserved miRNAs and 14 novel miRNAs after the removal of those that had already been found in strain 9509. When consolidated with the results from strain LT5a and mapped back to the strain 7498 genome, these two showed a strong degree of overlap resulting in 63 conserved and 45 novel miRNAs. These miRNAs were then further judged by the stringent criteria for plant miRNA annotation by sRNA-seq indicating that only 30 were highly confident based on structure and read count (Axtell and Meyers 2018). These miRNAs were then used to predict 163 targets with a psRNATarget score better than 2 (Dai and Zhao 2011), with roughly half corresponding to novel miRNAs.

The first prediction of miRNAs based on genome sequence and hairpin structure saw 413 possible miRNAs, and this number dropped to 58 and 59 once the miRNAs were being predicted based on sequencing results (Table 16.1). Of the 413 miRNAs, many were from recently discovered families, with only 121 that corresponding to those 58 families sequenced in 7498 at 119 genomic loci. While numbers of miRNA loci within families mostly agree, the copy number of a few families based on expression data differs from the 7498 genome survey as shown in Table 16.2. Perhaps the 24 copies of miR156 include a number of unexpressed pseudogenes from duplication events. When the strain 7498 and 9509 conserved miRNA families were compared 20 overlapped, while two were only found in the 9509 genome, and the 7498 study included 11 less commonly conserved one-member miRNA families not observed in strain 9509. This overlap of family and sequence number of highly conserved families suggests we have robust identification of the expressed, heavily conserved miRNA families, while lower confidence previously reported and novel miRNAs require further investigation to characterize.

While much attention is always paid to proper identification and mapping of miRNAs in the first sequencing experiments of a genome, measuring miRNA abundance is also essential. Since miRNA families have high sequence homology and target the same family of gene targets, these results are grouped by expression of certain

families. The three experiments studied strains LT5a, 9509, and 7498 which originated in China, USA, and Germany, providing a global perspective of the species. The control conditions were largely similar using Schenk & Hildebrandt medium at a pH of 5.8, with the known variations mainly being the 15 °C night time temperature, and relatively young cultures for LT5a and harvesting based on water surface coverage in 7498. While these expression results from strains across the world grown in control conditions vary in rank and abundance of miRNA families (Table 16.3), the same six families are within the top 5 in two of the three experiments demonstrating their prominent roles. As seen in Table 16.4, these miRNA families and their target gene families regulate growth, meristem development, and stress responses.

Strain 9509 was also exposed to 1uM ABA, which was shown to induce turion production irreversibly after 3 days (Wang et al. 2014b; Kuehdorf et al. 2014). At the 10 h time point, this hormonal stimuli changed the expression of 12 conserved miRNAs (over 100 reads in control, over twofold expression change in ABA), with the 169 and 396 families being underexpressed and the 159 and 168 families doubling in abundance (Michael et al. 2017). Then, at the 5-day time point, there were 28 miRNAs and targets with significant overexpression of the miRNA and underexpression of the mRNA compared to control and vice versa. Twelve of the miRNAs were novel miRNAs with relatively low expression, large fold change differences, and a wide variety of targets. Similar to the transcriptomic study at day 3, this experiment saw a decline in chloroplast proteins and an increase in polyphenol producing enzymes (Wang et al. 2014b; Michael et al. 2017).

The survey of miRNAs in strain 7498 in the control, cold, heat, abscisic acid, copper, kinetin, nitrate, and sucrose stimuli yielded a striking

Table 16.2 Copy number variation of miRNA families between three publications

miRNA family	7498 genome survey	9509 sRNA-seq	7498 sRNA-seq
156	24	6	9
159	1	3	4
169	9	5	7
396	11	5	9

Table 16.3 miRNA expression of control conditions of three strains of *Spirodela polyrhiza*

LT5a	7498	9509
156 (47%)	156 (41%)	160 (68%)
166 (24%)	168 (18%)	169 (7%)
167 (20%)	396 (16%)	166 (6%)
168 (5%)	169 (6%)	528 (5%)
169 (1%)	166 (4%)	159 (3%)

Table 16.4 Biological roles of prominent miRNA families

miRNA family	mRNA target family	Biological role
miR156	SPLs	Maintains juvenile tissues
miR166	HDZipIIIs	Regulates meristems
miR167	IARs	Reduced by drought
miR168	Ago1	Viral defense
miR169	NFYs	Drought and stress response
miR396	GRFs	Regulates meristems

result in that miR169c was between 33 and 82% of the reads in each condition, with large variability between the three biological replicates. This result was believed by the authors of the study to be an experimental artifact due to the lack of this expression in the other experiments, the only partial replication of the expression in the qPCR follow-up, and the current reputation of the SOLiD5500 sequencer. With this one sequence ignored and the dataset renormalized, we can accurately see the responses of other miRNAs to the various conditions. There were large increases in miR166 expression under the influence of cold and kinetin and miR168 in the heat and sucrose conditions. The meristem regulating 396 familiesy doubled expression in response to the heat, ABA, and copper stimuli. Finally, miR156, which maintains the juvenile, neotenous life cycle of the duckweed family, decreased over fourfold in response to sucrose, which was the condition responsible for 13 of the 19 instances of differential miRNA expression indicating that the mixotrophic lifestyle often used in laboratory experiments is quite different from duckweed grown in an outdoor setting.

Accurate miRNA annotation is quite difficult, since miRNAs are vastly outnumbered by similarly sized siRNAs in the genome, and even the more stringent miRNA prediction programs supply tens or hundreds of false predictions. An analysis in 2014 suggested that 75% of the land plant miRNA families in miRBase are questionable, especially those with only a single member (Taylor et al. 2014). In an attempt to manage the large number of submissions and false positives coming in, miRBase has established criteria for its high confidence miRNAs that analyze the structure of the hairpin, the read distribution along it, and the miR, and miR* read count. For plants in miRBase release 21, there are currently 6942 hairpins in 2408 distinct miRNA families, with only 587 from 227 families (9.7%) making the high confidence cutoff (Griffiths-Jones 2006). As an attempt to preserve miRNA annotation confidence, 21 of the leading minds of the field wrote the plant miRNA annotation criteria in 2008 that has since been updated by two of them thanks to new information and sequencing

capabilities (Meyers et al. 2008; Axtell and Meyers 2018). The plant miRNA annotation criteria are generally more stringent than the high confidence criteria, except for the latter's requirement of 10 miR* reads, since plant miRNA biogenesis is quite specific. Both miRNA studies in *Spirodela* annotated miRNAs based on homology according to the 2008 criteria, with most of these being well-conserved, high-confidence miRNA families. The conserved miRNAs with family names above 535 are relatively likely to be based off of lower confidence annotations in previous reports. The novel miRNAs from strain 9509 were predicted in 2017 using cutoffs very similar to the 2018 criteria, demonstrating a high degree of confidence, while those predicted in the LT5a and 7498 study had a lower degree of confidence. These authors reviewed all their data, with the revised criteria finding that 30 of the 47 hairpin structures met the current standards.

In addition to applying the stringent structural and read distribution filters above, the authors of the 2018 study verified miRNAs through a method called degradome sequencing where uncapped mRNAs are sequenced and aligned to miRNA target sites to measure evidence of precise miRNA cleavage above random mRNA degradation. There were several methods available at the time, and the authors chose the GMUCT2.0 library for its read length and minimal PCR amplification and the sPARTA program for its accuracy in analysis of the degradome data (Kakrana et al. 2014; Willmann et al. 2014). Biological triplicate libraries of the same eight conditions observed in the miRNA-sequencing study were sequenced on the Illumina NextSeq 500, yielding 911 million total reads. When running the sPARTA program, the *Spirodela* 7498 gene models were extended 150nt upstream and 250nt downstream, since many of the degradome reads were from the UTRs of the mRNAs. The sequencing verified activity of 66 miRNAs on 149 targets. For the 42 conserved miRNAs, the targets were mainly the transcription factor families reported in other plant species. While these essential developmental transcription factors mostly made up the

targets sequenced in over half of the conditions, 71% of the cleaved targets were sequence specific underscoring the importance of sampling a variety of post-transcriptional responses. Notably sucrose had the largest number of condition-specific results including metabolic and signaling proteins indicating a large shift in the mixotrophic lifestyle. This included a complete reversal where miR172 went from cleaving half as many targets as miR156 to twice as many despite being 0.4% of its expression. This suggested that sucrose may be inducing a less neotenous phenotype, and that highly expressed miRNAs are not necessarily highly active. Of the 81 novel miRNAs predicted within the three separate experiments of *Spirodela,* 24 were validated with 66 targets. This 30% validation rate, evenly spread between the three experiments, is consistent with similar surveys in other plant genomes thanks to the low expression and number of targets compared to conserved miRNAs, and the likelihood that novel miRNAs may be false predictions (Song et al. 2010; Li et al. 2010; Yang et al. 2013). While degradome evidence is a great way to confirm miRNAs, it does require co-expression and mRNA cleavage meaning that non-supported miRNAs may be found as active in later experiments with the right conditions and sequencing depth.

In order to provide other scientists easy access for further analysis, the raw data is available for LT5a results at GSE55208, 9509 at PRJNA308109, and 7498 at PRJNA473779 (SRP149336). As a second approach to increase transparency, ease replication, and enable further research, the data from the 2018 study and some of its analysis can be viewed in the Galaxy server as a history of the analysis, which includes the option of extracting the workflow and adapting it to analyze similar data Spirodela7498Galaxy-history (Afgan et al. 2016). Then, as a third method to make the data quick to review and useful to the community, the 7498 results are now displayed on an interactive viewer hosted by the Myers lab at the Danforth center https://mpss. danforthcenter.org/tools/mirna_apps/comPARE. php. Here the user can search for miRNAs, targets and sequences, see the expression across the 24 libraries, and download expression data (Fig. 16.1) (Nakano et al. 2006). The goal of this data accessibility was to enable other scientists to explore beyond the miRNAs, to the phased small interfering RNAs, the possible lncRNA intergenic targets in the degradome sequencing, or any other striking discovery within the datasets.

While the primary focus of both sRNA-seq experiments was to analyze miRNAs, Professor Jie Tang working with strain LT5a noted a surprising lack of 24nt RNAs typically found in plant genomes. These are often comparable in expression to the 22 and 21nt miRNAs, but they were rare as 7.3% of the small RNAs in strain
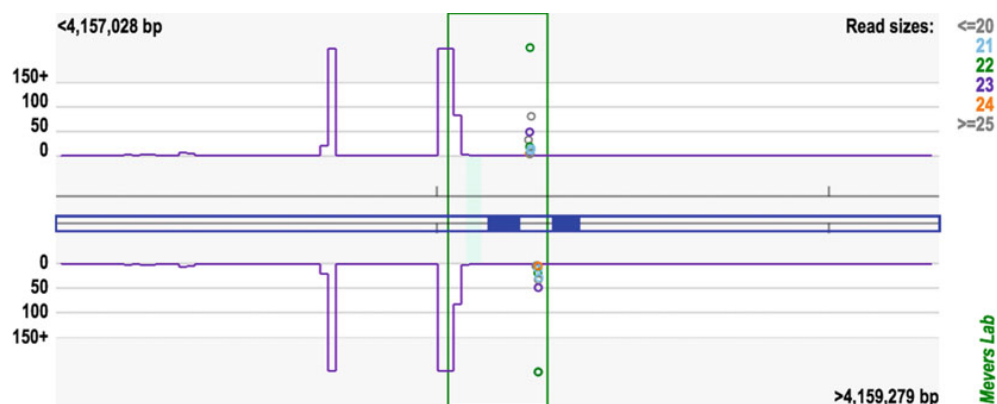
Fig. 16.1 View of small RNA browser showing high expression of the 22nt miR396d in the intron of the unknown protein Spipo10G0052600 in the control 1 library

LT5a, and 1% in 7498. In other plant species, 24nt RNAs are a part of the RNA-directed DNA methylation pathway where transposons are transcribed into single-stranded and then double-stranded RNA, diced into 24nt heterochromatic small RNAs, and then used to guide a protein network that methylates matching sequences and then silences them as heterochromatin. Accordingly, Michael et al. also studied DNA methylation in the *Spirodela* 9509 genome and found it to be the least methylated plant sequenced! This DNA methylation pathway in duckweeds is a new and exciting field of study summarized in Chap. 5 that appears to be the cutting edge of small RNA research in the Lemnaceae.

References

Afgan E, Baker D, van den Beek M et al (2016) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44:W3–W10. https://doi.org/10.1093/nar/gkw343

An J, Lai J, Sajjanhar A et al (2014) miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. BMC Bioinform 15:275. https://doi.org/10.1186/1471-2105-15-275

Axtell MJ, Meyers BC (2018) Revisiting criteria for plant miRNA annotation in the era of big data. Plant Cell 30:tpc.00851.2017.   https://doi.org/10.1105/tpc.17.00851

Brannan CI, Dees EC, Ingram RS, Tilghman SM (1990) The product of the H19 gene may function as an RNA. Mol Cell Biol 10:28–36. https://doi.org/10.1128/MCB.10.1.28

Dai X, Zhao PX (2011) PsRNATarget: A plant small RNA target analysis server. Nucleic Acids Res 39: W155–W159. https://doi.org/10.1093/nar/gkr319

Denman RB (1993) Using RNAFOLD to predict the activity of small catalytic RNAs. Biotechniques 15:1090–1095

Fahlgren N, Carrington JC (2010) miRNA target prediction in plants. Humana Press, pp 51–57

Franco-Zorrilla JM, Valli A, Todesco M et al (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet 398 (39):1033. https://doi.org/10.1038/ng2079

Griffiths-Jones SM (2006) The microRNA sequence database. Methods Mol Biol 342:129–138

Howell MD, Fahlgren N, Chapman EJ et al (2007) Genome-wide analysis of the RNA-dependent RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. Plant Cell 19:926–942. https://doi.org/10.1105/tpc.107.050062

Johnson C, Kasprzewska A, Tennessen K et al (2009) Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. Genome Res 19:1429–1440. https://doi.org/10.1101/gr.089854.108

Jones-Rhoades MW, Bartel DP (2004) Computational identification of plant MicroRNAs and their targets, including a stress-induced miRNA. Mol Cell 14:787–799. https://doi.org/10.1016/J.MOLCEL.2004.05.027

Kakrana A, Hammond R, Patel P et al (2014) SPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. Nucleic Acids Res 42:1–13. https://doi.org/10.1093/nar/gku693

Kuehdorf K, Jetschke G, Ballani L, Appenroth K-J (2014) The clonal dependence of turion formation in the duckweed *Spirodela polyrhiza*—an ecogeographical approach. Physiol Plant 150:46–54. https://doi.org/10.1111/ppl.12065

Li Y-F, Zheng Y, Addo-Quaye C et al (2010) Transcriptome-wide identification of microRNA targets in rice. Plant J 62:742–759. https://doi.org/10.1111/j.1365-313X.2010.04187.x

Meyers BC, Axtell MJ, Bartel B et al (2008) Criteria for annotation of plant MicroRNAs. Plant Cell 20:3186–3190. https://doi.org/10.1105/tpc.108.064311

Michael TP, Bryant D, Gutierrez R et al (2017) Comprehensive definition of genome features in *Spirodela polyrhiza* by high-depth physical mapping and short-read DNA sequencing strategies. Plant J 89:617–635. https://doi.org/10.1111/tpj.13400

Nakano M, Nobuta K, Vemaraju K et al (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res 34:D731–D735. https://doi.org/10.1093/nar/gkj077

Scott MS, Ono M (2011) From snoRNA to miRNA: dual function regulatory non-coding RNAs. Biochimie 93:1987–1992.   https://doi.org/10.1016/J.BIOCHI.2011.05.026

Song C, Wang C, Zhang C et al (2010) Deep sequencing discovery of novel and conserved microRNAs in trifoliate orange (citrus trifoliata). BMC Genom 11:431. https://doi.org/10.1186/1471-2164-11-431

Taylor RS, Tarver JE, Hiscock SJ, Donoghue PCJ (2014) Evolutionary history of plant microRNAs. Trends Plant Sci 19:175–182. https://doi.org/10.1016/J.TPLANTS.2013.11.008

Wang W, Haberer G, Gundlach H et al (2014a) The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun 5:1–13. https://doi.org/10.1038/ncomms4311

Wang W, Wu Y, Messing J (2014b) RNA-Seq transcriptome analysis of Spirodela dormancy without reproduction. BMC Genom 15:60. https://doi.org/10.1186/1471-2164-15-60

Willmann MR, Berkowitz ND, Gregory BD (2014) Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes-GMUCT 2.0. Methods 67:64–73. https://doi.org/10.1016/j.ymeth.2013.07.003

Yang J, Liu X, Xu B et al (2013) Identification of miRNAs and their targets using high-throughput sequencing and degradome analysis in cytoplasmic male-sterile and its maintainer fertile lines of brassica juncea. BMC Genom 14:9. https://doi.org/10.1186/1471-2164-14-9

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415. https://doi.org/10.1093/nar/gkg595