Computational prediction of candidate microRNAs and their targets from the completed *Linum ussitatissimum* genome and EST Database

Tiffanie Y. Moss, Christopher A. Cullis

Department of Biology, Case Western Reserve University, Cleveland, Ohio, USA

Abstract

Linum usitatissimum (Flax) is an important agronomic crop grown for its fiber (linen) and oil (linseed oil). In spite of thousands of years of breeding some fiber varieties have been shown to rapidly respond to environmental stress with heritable changes to its genome. Many microRNAs (miRNAs) appear to be induced by abiotic or biotic conditions experienced through the plant life cycle. Here a bioinformatics approach is used to screen for miRNAs previously identified in other plant species, as well as to predict putative miRNAs unique to a particular species which may not have been identified as they are less abundant or dependent upon a specific set of environmental conditions. Twelve miRNA genes were identified in flax on the basis of unique pre-miRNA positions with structural homology to plant pre-miRNAs and complete sequence homology to published plant miRNAs. These miRNAs were found to belong to seven miRNA families, with an additional two matches corresponding to as yet unnamed poplar miRNAs and a parologous miRNA with partial sequence homology to mtr-miR4414b. An additional 649 novel and distinct flax miRNA genes were identified to form from canonical hairpin structures and to have putative targets among the ~30,000 flax Unigenes.

Introduction

Linum usitatissimum (Flax) is an important agronomic crop grown for both its fiber (linen) and oil (linseed oil). The two different uses have required breeding to develop cultivars either with higher oil content and seed yield or reduced branching and increased height. In spite of thousands of years of breeding some fiber varieties have been shown to rapidly respond to environmental stress with heritable changes to its genome, as well as changes in phenotype over the course of a single generation.[1-5] The inducibility of rapid structural variation in the genome makes flax a unique plant model for identifying the mechanisms controlling genome stability. The variation observed in the rapid response includes changes in genome size, copy numbers of ribosomal DNA, and structural rearrangements.[3,5,6] Computational microRNA (miRNA) analysis of the flax genome provides a foundation for subsequent research on miRNA function in flax. Therefore an analysis of miRNAs and their targets is expected to identify unique flax miRNAs not previously identified in other plant genomes.

In plants, endogenous miRNAs are known to be involved in many biological and metabolic processes, including plant architecture, meristem development, stress response and signal transduction.[7,8] Many miRNAs appear to be induced by abiotic or biotic conditions experienced through the plant life cycle.[9-11] A bioinformatics approach is a way to screen for miRNAs previously identified in other plant species, as well as to predict putative miRNAs unique to a particular species which may not have been identified as they are less abundant or dependent upon a specific set of environmental conditions. MiRNAs are 19-21nt non-coding RNA molecules derived from hairpin forming pre-cursor sequences and their biogenesis involves multiple enzymes over many steps.[12] Pri-miRNAs, the primary transcripts, are transcribed from the genome and cleaved into pre-miRNAs which form the characteristic hairpin structure processed by a Dicer-like enzyme.[13] The mature miRNA sequence is cleaved from the precursor while still bound to its complement, known as miRNA*, in a double-stranded conformation (miRNA:miRNA*). This complex is then bound by the Argonaute protein and the complementary strand is released before the guide strand is incorporated into the RNA Induced Silencing Complex to enable targeted mRNA repression or degradation prompted by mRNA cleavage.[14]

MiRNAs are just one of the three main classes of small RNAs: small interfering RNAs (siRNAs), microRNAs and Piwi-interacting RNAs (piRNAs). Although plants have not yet been shown to produce piwiRNAs, many of their siRNAs have developed similar regulatory roles to those of piwiRNAs in other organisms.[14] While siRNAs originate from double-stranded RNA and miRNAs are encoded as part of the transcriptome, the biogenesis of siRNAs and miRNAs have similar stages during their biogenesis in plants, including their slicing activity against their targets.[15] This can make it difficult to distinguish miRNAs by their mature sequence or targets alone. Furthermore, Wang *et al.* have shown that there is little pre-miRNA sequence conservation between species; thus, conventional sequence-alignment based methods are insufficient for identifying anything other than very close miRNA homologs on the basis of pre-miRNA sequence conservation.[16]

Correspondence: Tiffanie Yael Moss, Case Western Reserve University, Department of Biology, 2080 Adelbert Road, Millis 127, Cleveland, Ohio 44106-7080, USA.
Tel. +1.216.682.5105 - Fax: +1.216.368.4672.
E-mail: tiffanie.moss@case.edu

Key words: microRNA, bioinformatics, genome, flax, EST.

Acknowledgements: we are very grateful for the assistance of Kyle Logue and the rest of David Serre's team at the Cleveland Clinic Foundation for their assistance with the Perl scripts used in parsing this data.

Contributions: TYM did the analysis, wrote the Perl Scripts and was responsible for most of the writing of the manuscript; CAC contributed to the design of the study, and to the writing of the manuscript.

Conflict of interests: the authors report no potential conflict of interests.

Received for publication: 7 February 2012.
Revision received: 11 April 2012.
Accepted for publication: 17 April 2012.

Journal of Nucleic Acids Investigation 2012; 3:e2
doi:10.4081/jnai.2012.e2

Guidelines for miRNA classification in plants have been outlined by Ambros *et al.* and Meyers *et al.* requiring that the miRNAs be derived from a stem-loop precursor in which *i) The miRNA and miRNA* are derived from opposite stem-arms such that they form a duplex with two nucleotide, 3 ' overhangs; ii) base-pairing between the miRNA and the other arm of the hairpin, which includes the miRNA*, is extensive such that there are typically four or fewer mismatched miRNA bases; and iii) asymmetric bulges are minimal in size (one or two bases) and frequency (typically one or less), especially within the miRNA/miRNA* duplex.*[17,18] According to Meyers *et al.*, computational analysis indicating preservation of stem-loop precursor embedded with the mature miRNA sequence provides strong support for miRNA annotation.[18] Thus, computational analysis is a rational first step in the analysis of miRNAs in a newly sequenced genome.

Bioinformatics analysis of a genome or its EST Database is often the first step in a small RNA study due to its minimal costs and ability to identify miRNAs. The putative miRNAs can also be examined for position effects within

the genome itself and potential EST sequences they may be associated with either as targets or precursors. Preliminary characterization provides a framework for future analysis of miRNA genes and their roles in key traits.

An *ab initio* approach to identifying potential miRNAs examines genomic and/or UNIGENE sequences for structural similarities and characteristics in keeping with miRNA biogenesis. Algorithms designed to examine potential precursor sequences for their ability to fold back into characteristic hairpin sequences with minimal free-energy values and appropriately placed mature sequences within the stem with few gaps and mismatches between the miRNA and its miRNA* are utilized. This step is computationally intensive as each sequence is examined for its ability to form hairpin structures and the associated free energy calculation. A comparative genomics approach identifies miRNAs based on sequence conservation between species. It has been used in cotton, soybean, wheat, potato, apple, switchgrass and citrus to identify conserved mature miRNAs and then to expand out from those sequences to determine if a canonical hairpin can be formed.[8,19-24] However, conserved and *de novo* miRNAs can be identified in the same workflow (Figure 1) by combining these approaches to first determine if the canonical hairpins which meet the criteria outlined by Ambros *et al.* are able to form, and then scanning the predicted hairpin and their predicted mature miRNAs for sequence conservation with plant miRNAs available in the public datasets.[18]

The recent availability of the genome sequence from the oilseed flax variety Bethune has allowed for an initial examination of flax's potential miRNA profile. In this study the flax genome, Bethune v.09, and the flax seed EST Database, UNIGENE, were used to predict conserved and novel miRNAs by way of the BLASTn and *De Novo Prediction of MicroRNA-Coding Regions in a Single Plant-Genome* (NovoMIR) programs. NovoMIR has been previously shown to have a high sensitivity and specificity for other plant genomes, however this will be the first test of novoMIR on the flax genome.[25] Computational prediction of novel and conserved miRNAs using both the genome and EST database provides an alternative solution to the difficulties associated with large-scale experimental validation of miRNA expression such as cost and limited tissue or conditionally expressed microRNAs. The predictions makes expression data associated with these ESTs available for future experiments examining functional characteristics of miRNAs in flax.[26] Here we report the identification of conserved and novel miRNAs and examined potential targets of these miRNAs.

Materials and Methods

Sequences and databases

The flax database of 30,640 full-length flax ESTs collected from seed libraries known as UNIGENE was provided by Raju Datla in June 2011 and are now available.[26] Unigenes were used in the prediction of miRNAs as well as target sequences.

The Flax genome v.09 is comprised of 302 Mb of non-redundant sequence representing an estimated ~85% genome coverage at 95X Illumina coverage and the database of predicted gene regions of the genome was made available by Michael Deyholos (University of Alberta) prior to publication. The genome is publicly available at www.linum.ca.

To identify all potentially conserved miRNAs in the flax genome, all mature and hairpin miRNA sequences were downloaded from the miRBase sequence database, release 18.0, on December 11, 2011 and all known mature miRNA sequences were downloaded from the Plant MiRNA Database (PMRD) on November 15, 2011.[27-29] From miRBase v.18 4677 mature plant miRNA sequences were extracted and then were concatenated with the 10,121 mature miRNAs in the PMRD and redundant matches excluded. This provided a dataset of 7369 mature miRNAs and 4014 pre-miRNAs.

Software

NovoMIR was used to predict potential *de novo* miRNAs and was downloaded from http://www.biophys.uni-duesseldorf.de/~teune/.[25] NovoMIR finds all possible sequences capable of forming the canonical hairpin structure consistent with miRNA biogenesis. These sequences were then examined for sequence similarity to miRNA precursor sequences catalogued in miRBase and PMRD. The Vienna software package 1.8.4 containing RNALfold and the program RNAshapes,[30-32] are required by novoMIR and were also downloaded from the web site as well as the RNAFold module used to test the miRTour predicted structures.

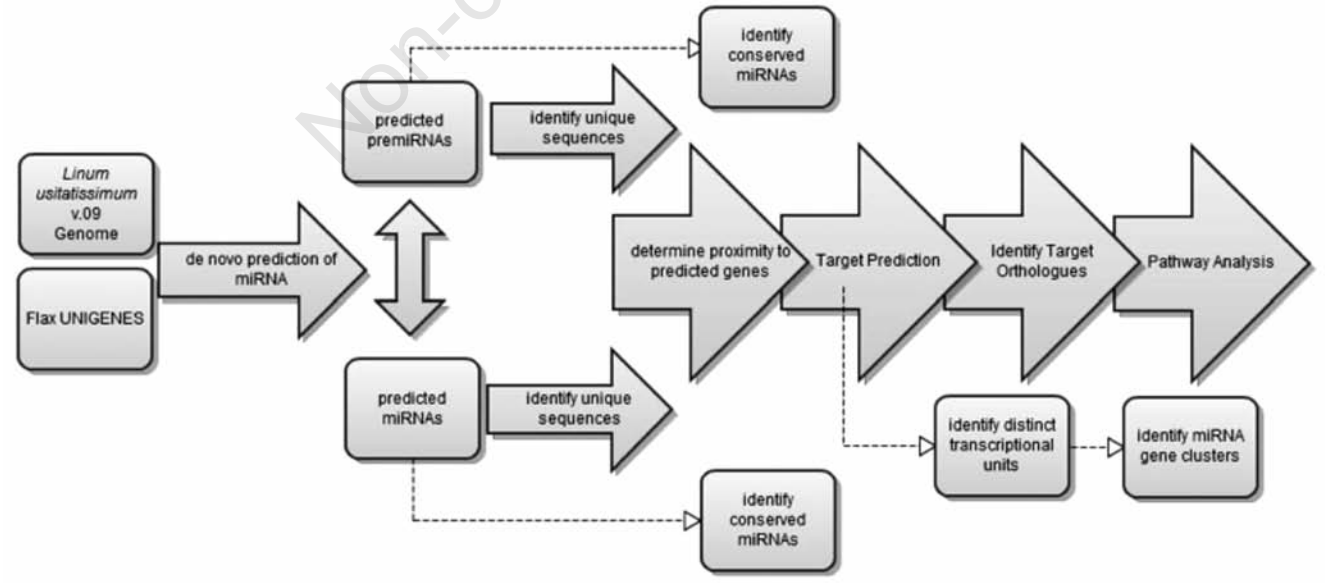

Figure 1. Flowchart-method for bioinformatic prediction of flax microRNAs and analysis of their targets. Following prediction of mature and precursor sequences by novoMIR, a comparative search of the sequences was run against plant entries in miRBase and the Plant microRNA database to identify conserved miRNAs.

Visualization Applet for RNA secondary structure (VARNA) is a Java program which was used to draw and annotate structural representation of the structures predicted by the Vienna software[33] BLASTN 2.2.24 release (23 August 2010) was used to identify conserved pre-miRNA sequences among predicted pre-miRNAs using the miRBase v.18 dataset.[34] BLASTn was used to identify conserved mature miRNA sequences in the flax genome relying on the plant miRNA dataset in miRBase v.18 and the PMRD as a reference set. Perl scripts were developed and used in parsing the data.

Identification of microRNAs in the flax genome and in the EST database

The algorithm novoMIR was used in an ab initio approach to identify putative miRNAs and their precursors based on structural characteristics consistent with miRNA biogenesis. novoMIR was designed and tested on plant genomes. The prediction of miRNAs does not rely on knowledge of targets nor does it use comparative genomics. Rather, novoMIR is a Perl script that uses a series of filters and statistical models to discriminate a pre-miRNA from all other RNAs and to locate the miRNA:miRNA* complex in a putative pre-miRNA. novoMIR relies on RNAfold and RNAshapes for secondary structure prediction of precursors and the associated free energy measurements.[25]

The Bethune genome and Unigene sequences were input into novoMIR. All non-miRNA hairpin sequences were removed and the remaining putative miRNAs and their hairpins were compiled into a database. NovoMIR provides a maximum of six candidate mature miRNA sequences for each pre-miRNA, however this does not exhaustively identify all possible miRNAs.[25] The hairpin and mature miRNA sequences were then compared with the miRBase and PMRD databases, using Blastn, to identify conserved miRNAs with sequence homology to known miRNAs. The matches were identified as conserved miRNAs if the mature sequences were exact matches or had <2 mismatches to those in the databases and the mismatches did not occur in the region between nt 9 and 12. A Perl script was written to classify each putative miRNA gene as being located on a scaffold or contig without any predicted genes, or if the miRNA was inside, outside or overlapping predicted genes.

The miRTour program

The miRTour program was recently published and was used as an additional test of the Unigene dataset.[35] The Unigene database was entered into miRTour to identify conserved miRNAs and their targets. miRTour excludes protein coding ESTs and examines the non-coding ESTs for hairpins to form pre-miRNAs. It then looks for conserved mature miRNAs within hairpins with <2 mismatches miRNA:miRNA* and tries to identify a target for the miRNA within the same uploaded EST database.

Target prediction

The Plant Small RNA Analysis Server (psRNATarget) provides reverse complementary matching between miRNAs (and ta-siRNAs) and target transcript and determines target site accessibility by calculating unpaired energy (UPE) necessary for opening the secondary structure around the miRNAs target site while distinguishing between translational and post-transcriptional inhibition.[36] All mature miRNA sequences derived from the flax genome and Unigenes were entered into psRNATarget at http://plantgrn.noble.org/psRNATarget/ with Unigenes used as the transcript dataset for the target search.[36] PsRNATarget is able to distinguish between translational and post-transcriptional inhibition according to the binding pattern seen between positions 9-11. PsRNATarget was run with a maxExpect of 3, a length of 20 for complementarity scoring, 25.0 maximum UPE to unpair the target site, a flanking length of 17 bp upstream and 13 bp downstream of the target site to be used for target accessibility analysis and nt 9-11 used as the range of central mismatch leading to translational inhibition. Unigenes which were identified as targets, but which also had sequence homology to predicted genes from which the miRNAs were derived, were excluded from the dataset as being a miRNA with a predicted target.

Target annotation and pathway enrichment analysis

Targets identified by psRNATarget were curated through the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (KAAS) at http://www.genome.jp to identify putative function (via BLASTx against non-redundant sequences) and assign them to a biological network according to their orthology.[37,38] A bidirectional best hit assignment method was used against the plant genomes of: *Arabidopsis thaliana* (thale cress), *Arabidopsis lyrata* (lyrate rockcress), *Oryza sativa japonica* (Japanese rice), *Ostreococcus lucimarinus, Ostreococcus tauri*, and *Cyanidioschyzon merolae*.[38] KEGG BRITE was then used to identify functional hierarchies among the miRNA Targets and assign them to a biological process, cellular component or molecular function.[39]

Results

Identification of potential microRNAs in flax

The novoMIR program identified 2324 potential pre-miRNAs and 14,405 potential mature miRNAs from the flax genome Bethune v.09 which contained 56,344 scaffolds or contigs derived from the 302 Mb of non-redundant sequence representing ~85% genome coverage (Table 1). Hairpin structures with the miRNA position indicated are available for all predicted miRNAs cited in this work in Supplementary Files 1 and 2. Genome and Unigene derived structures which did not meet the current criteria for miRNA biogenesis are available in in Supplementary Files 3 and 4 respectively. With removal of redundant matches, 3450 unique miRNAs and 1632 pre-miRNAs were obtained. Among the 30,640 Unigenes, 10,692 mature miRNA sequences of which 2688 were unique were identified and 1954 precursor sequences of which 1496 were unique were identified.

The Unigene and genome derived pre-miRNAs were compared and no redundant full length matches were identified. However, in a BLAST search of the genome, 507 of the Unigene pre-miRNAs could be aligned completely. The small overlap is likely due to the nature of the novoMIR algorithm which filters

Table 1. Flax micro RNA gene summary based on computational analysis of the Bethune genome

Summary	Genome v.0910	Unigene
MiRNAs - total	14405	10692
Pre-miRNAs - total	2324	1954
Unique miRNA sequences	3450	2688
Unique pre-miRNA sequences	1632	1496
Conserved miRNA genes	2	8
Conserved miRNA families	2	6
Distinct miRNA gene units (known to have targets)	220	434
Clusters of miRNA genes (known to have targets)	8	0
Total miRNA genes in flax with identified targets	649	

the five subsequences with the best locally stable structures derived from 1000 nt placed through RNALFold in order to expedite the first step of the algorithm.[25] When genomic sequences are put through novoMIR they are subdivided into 1000 nt fragments overlapping by 400 nt.[25] It is not unlikely that there would be many cases in which the genomic region associated with a Unigene would not be completely within the 1000 nt window of the algorithm as it scrolls the region of the genome from which the Unigene originates. Therefore it is not unlikely that the five best structures from a given region would not correspond with the five best structures from the Unigene of that region as the 1000 nucleotides examined for each may be off by at least 400 nt.

The novoMIR predicted miRNAs were classified to miRNA families on the basis of sequence conservation to plant miRNAs listed in the PMRD and miRBase databases using a BLASTn algorithm with thresholds of 4E-06 (>21nt and <2 mismatches). Five miRNA genes from four families were identified among the Unigene dataset: miR159, miR165b, miR167, miR319 and unnamed miRNAs: PMRD-ptc-miRf10271-akr and PMRD-ptc-miRf10178-akr. Interestingly, we also had two matches to aly-miR167b* (which is antisense). Using the miRTour algorithm we identified an additional five miRNA genes belonging to three families: miR164, miR167 and miR408. MiR167 was identified as a conserved miRNA gene in the unigenes: LUSHE1NG-RP-072_C09_X, Contig6197, and LUSTE1AD-RP-276_D15_X. The hairpin associated with the miRNA derived from Contig6197 is shown in Figure 2.

In the flax genome, two miRNAs with conserved sequences to those listed in the database could be found among the predicted pre-miRNAs. Precursors miR154 and miR791 in the flax miRNA dataset (Supplementary File 1) were found to have sequence conservation (20-21nt match with <2 mismatch) with other plant miRNAs belonging to miRNA families 169 and 319 respectively. In total, we were able to identify 12 conserved miRNA genes in flax belonging to seven miRNA families with an additional two matches corresponding to as yet unnamed poplar miRNAs (Table 2) with miR167 being the most abundant conserved family. According to miRBase (Release 18), miR167 has been associated with 25 different plant species. In *Vitis vinifera* it was found that the family miR167 had differential expression pattern according to tissue type, as did the families miR166, miR169 and miR171 which is consistent with additional findings in maize.[40-42] The miR167 family is also thought to be involved in auxin regulation, which is important for many diverse developmental responses, including cell elongation, division and differentiation.[43]

All the conserved miRNAs identified in the

Table 2. Conserved microRNAs in flax - 21 or 22 nt with <2 mismatches to plant miRNAs published in miRBase or plant MiRNA database.

miRNA family	Genome v.0910	Unigene	Totals
159	0	2	2
164	0	1	1
165	0	1	1
167	0	3	3
169	1	1	2
319	1	1	2
408	0	1	1
Total	2	10	12

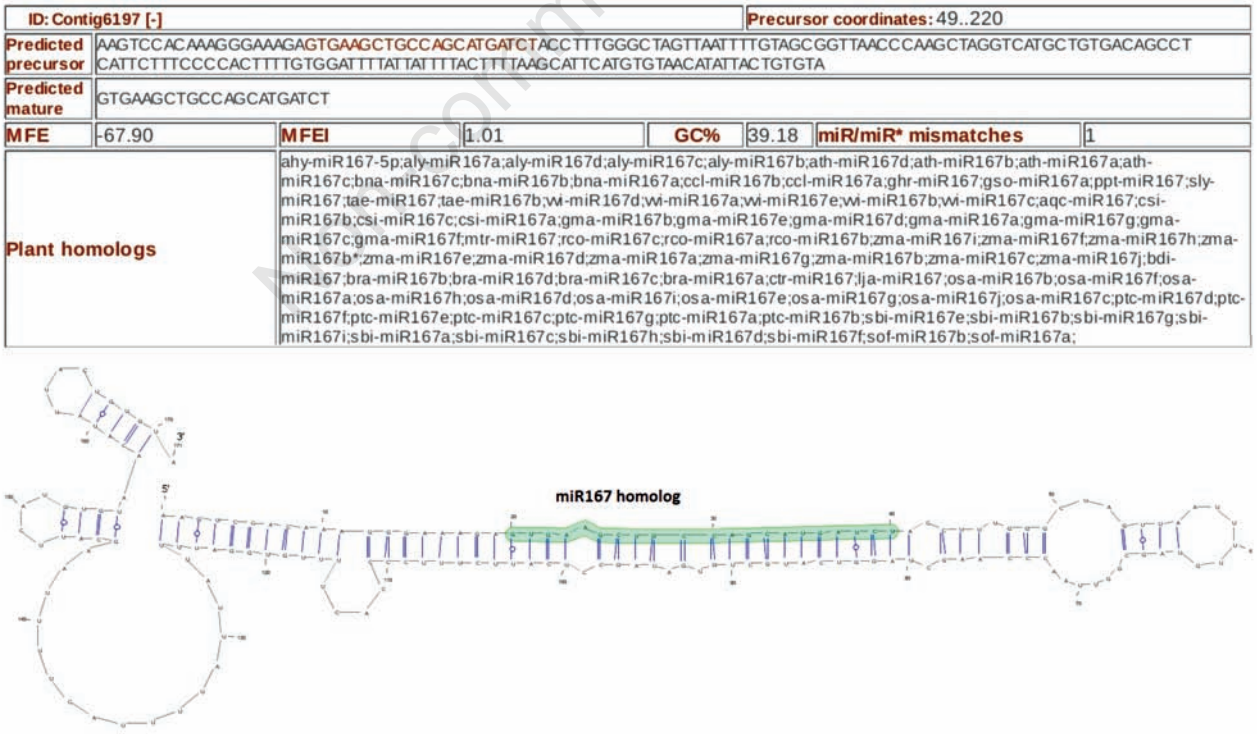

Figure 2. MicroRNA hairpin predicted from Flax Unigene Contig6197 derived using miRTour. The 171 nt hairpin pre-miRNA was folded using the RNAFold module in the Vienna package and the microRNA region highlighted using Visualization Applet for RNA secondary structure (VARNA).[33]

genome analysis were derived from matches between other plant miRNAs and the predicted flax miRNA precursors. Although the predicted flax mature miRNA sequences associated with these miRNA precursors did not harbor any exact matches, the precursors themselves were also found to have sequence homology to ptc-MIR169i and ptc-MIR319f with e-values of 1e-12 and 5e-19 respectively. As poplar is considered a close relative to flax, it is reasonable to see sequence similarity even among the precursor sequences. According to Thakur *et al.*, positional conservation in plant miRNAs is observed in two distinct blocks, 2-13 and 16-19 with the conservation at position 2-12 found to be adequate.[44] Therefore, it could be argued that several additional conserved miRNAs have been identified, including a match to mtr-miR4414b, with a perfect match at positions 1-17. MiR4414b is thought to be involved in arbuscular mycorrhizal symbiosis formed by most flowering plants.[45]

All putative miRNA genes derived from the genome were categorized as one of the following categories, overlapping+outside predicted genic regions, inside predicted genic regions or as being located on scaffolds for which no genes are predicted (often scaffolds <1Kb). We identified a total of 134 putative pre-miRNA sequences which overlap or are located outside genic regions (97 outside and 37 overlap), of which 74 were unique. Fifty-two pre-miRNA sequences were located inside predicted genes, of which 48 were unique. Within scaffolds and contigs for which no genes had been predicted (empty scaffolds) 2141 putative pre-miRNAs were identified, of which 1510 were unique. Although the flax genome is still undergoing annotation, we consider these empty scaffolds as intergenic regions. Therefore a total of 1584 unique putative pre-miRNAs were considered intergenic regions and 48 were located inside genes. This is consistent with the assumption that most small RNAs originate from intergenic regions.[45] However, it should be noted that annotation programs used for analyzing RNAseq data often filter out reads mapped to coding sequences (CDS) and thereby miss potential miRNAs, even if the miRNA match is a chance homology to a CDS, resulting in the miRNA being unidentified.[44] In a recent miRNA study in corn, miRNAs belonging to miR160, miR166, and miR194 families were found to overlap with a CDS or were homologous to a CDS but were later confirmed by RACE.[44] For this reason, the 48 putative miRNAs derived from inside 15 different genes were also considered, but were filtered further according to their location within the gene. Of the 15 genes predicted to form miRNAs, four contained the precursor sequences in their introns or UTRs - g44356, g47829, g47848, and g45274 - and are the most likely candidates for being miRNA genes. Also likely are the two miRNAs found to overlap outside the gene - g47837and g47873. Three overlapped both intron and CDS regions - g21140, g4494, g47905 - and could be activated if facilitated by alternative splicing to produce only the miRNA, or the miRNA could be produced after splicing which would produce both the miRNA and spliced host mRNA.[46]

We initially determined which miRNA genes were predicted to have targets and then used this subset to identify whether they were likely to be independently transcribed or transcribed as a multi-miRNA cluster. MiRNA genes located in introns and intergenic regions were considered to be distinct gene units. This information was used to count the number of miRNA genes and to identify any clusters of miRNAs which are >100 nt apart but not separated by more than 10,000 nt. Although the flax genome is still undergoing annotation, we took the view that scaffolds or contigs empty of predicted genes (*empty scaffolds*) should be annotated as intergenic regions. Therefore we examined them for clusters of putative miRNAs as well. As our initial search for pre-miRNA sequences allowed for pre-miRNA sequences to overlap, we limited our count of total miRNA genes to those which were >100 nt from the start of another putative pre-miRNA and to those not located within a single predicted gene. This gave a total count of 220 miRNA genes in the flax genome with an additional 434 in the Unigene data (each unigene was considered to possess at most a single transcribable unit). Five pre-miRNAs were found to be redundant, existing in both the genome and unigene datasets, leaving a total of 649 distinct miRNA genes (Table 1). Although more predicted miRNAs were identified by novoMIR in the genome than in the Unigenes dataset as would be expected, more predicted miRNA genes were identified in the Unigene dataset than in the genome. This is likely due to the many predicted young miRNAs which were likely identified in the genome. Young miRNAs are thought to be transitory and often lack targets.[10] As our criteria limited predicted miRNAs to those with potential targets in the Unigene dataset, these were excluded from our dataset and the Unigene dataset appears enriched for predicted miRNAs using this method. As with all the predicted miRNAs discussed here, quantitative stem-loop RT-PCR could be used to validate the expression profile of these miRNAs within specific tissues.[47]

A miRNA cluster was defined as pre-miRNA sequences >100 nt from the next start site and <10000 nt from another pre-miRNA. These parameters identified eight clusters among the flax genome dataset of pre-miRNAs with targets. Clusters were found in scaffolds 639, 755, 2502, 3375, 3907, 4004, 4176, and 4264. All the clusters identified were found on scaffolds lacking predicted genes. Scaffolds 4004 and 755 each possess three miRNAs in their clusters distributed over ~1800 nt and ~1000 nt in length respectively. Each mature miRNA located within a cluster was most often found to target genes which differed from those of other miRNAs in the same cluster. However, as nearly all the targets of miRNAs associated with clusters are of unknown function, we are unable to determine whether or not the gene regulation could be coordinated.

Prediction of flax microRNA targets

At high stringency, psRNATarget used the 30,640 Unigenes as a transcript dataset to identify 1170 targets for the novoMIR generated miRNAs derived from Unigenes (Supplementary File 5). Many sequences were targets of multiple miRNAs and 582 were unique target sequences. Among genome-derived sequences, 4248 unique miRNA-target interactions were identified. As several miRNA sequence possibilities are associated with a single precursor, the number of miRNA-target interactions was reduced to correspond with a single transcriptional unit for a total of 211 miRNA-gene and target interactions and 209 unique targets.

Of the unique 211 transcribed miRNA gene/target combinations among flax genome miRNAs, we identified two genes which are targets for more than one miRNA. One of these genes was UNIGENE Contig6926 which is targeted by pre-miRNAs derived from g21140, which has an unknown function, and a pre-miRNA derived from g44356, which has sequence similarity to Populus trichocarpa RNA polymerase IV subunit. The miRNA:target duplex structures can be found in Supplementary File 6. BLASTx revealed that the putative target, UNIGENE Contig 6926, has sequence similarity with a Populus trichocarpa hypothetical protein of unknown function based on a search of non-redundant sequences.

A putative miRNA derived from g21140 was also found to target Contig 3977, as did the putative miRNA derived from g46976, a sequence with similarity to unnamed protein DUF677 with the PFAM domain PFAM: PF05055. BLASTx showed UNIGENE Contig 3977 has limited sequence similarity to a Populus trichocarpa predicted protein of unknown function. The miRNA:target duplex structures can be found in Supplementary File 7. As miRNAs predicted from the Unigenes listed here show no sequence similarity to known miRNAs they will need to be experimentally validated in order for their identity as species-specific miRNAs in flax to be confirmed.

Additionally, we found several cases where a miRNA was predicted to target more than one Unigene or instances where clusters were seen to regulate multiple genes. Using the

KAAS, the targets were annotated according their predicted orthology to other known plant genes. In scaffold1238, various miRNAs which could all be derived from the same pre-miRNA were found to target two different genes: the EIF4E, translation initiation factor 4E or RP-L29e, large ribosomal subunit 29e (Supplementary File 8). Another example can be seen in scaffold1389 where FAD8, Omega-3 fatty acid desaturase and Threonine synthase could be targeted by two different miRNAs derived from the same pre-miRNA (Supplementary File 9). In scaffold1966, EIF3D, translation initiation factor 3D or UTP18, U3 small nucleolar RNA-associated 18 protein can likewise be generated from a single precursor (Supplementary File 10). Lastly, in scaffold 2548, inside gene g45274, multiple mature miRNAs appear capable of being made with three different targets: cytochrome C oxidase assembly protein, light harvesting complex1 chloroplast, and PP1C - protein phosphatase (Supplementary File 11).

MicroRNA target annotation

All genome and Unigene derived miRNA targets were curated through KAAS to annotate targets according to sequence similarity with other plant genes using a bidirectional best hit assignment. KAAS assigned a Kegg Orthology (KO) number to each target capable of being classified. A total of 170 KO numbers were assigned to the flax miRNA targets. Of these, 16 KOs were assigned to more than one gene. A detailed list of the Unigenes targeted with the KO identifier and predicted function based on orthology is given in Supplementary File 12.

Target pathway analysis

Targets were assigned to the Network hierarchies of Metabolism, Genetic Information Processes, Environmental Information Processes, and Cellular Processes using the KEGG BRITE hierarchies. Within each hierarchy proteins belonging to several families were identified. The Metabolism hierarchy is associated with 12 protein families according to KEGG BRITE. 106 of the 170 flax miRNA target KEGG orthologs were assigned to nine of these families with the most notable being 73 different classes of enzymes and 13 different proteins associated with photosynthesis. Among the enzymes targeted, oxidoreductases, transferases and hydrolases were the most abundant in this dataset (Figure 3).

Six orthologs were shared between the Metabolism and Genetic Information Processing hierarchies: K01868 (threonyl-tRNA synthetase), K02726 (PSMA2; 20S proteasome subunit alpha 2), K03514 (REV1; DNA repair protein REV1), K06269 (PPP1C; protein phosphatase 1, catalytic subunit), K09580 (PDIA1, P4HB; protein disulfide-isomerase A1), and K15362 (BRIP1, BACH1, FANCJ; fanconi anemia group J protein). The last was disregarded as it is a protein unlikely to be associated with a biological pathway in plants. The Genetic Information Processing hierarchy contained 83 of the flax miRNA target KEGG orthologs and all 14 classes of proteins were represented. Thirty-two of these orthologs were specific to the class of proteins associated with ribosomes and an additional five to ribosome biogenesis. Among the Environmental Information Processing heirarchy, orthologs were assigned to two of the seven protein families and two of the 11 protein families within the Cellular Processes hierarchy (Figure 4). A more detailed breakdown of the genes within each of these major classes is given in Supplementary File 13.

Discussion

Twelve miRNA genes were identified in flax on the basis of unique pre-miRNA positions with structural homology to plant pre-miRNAs and complete sequence homology to published plant miRNAs, thereby meeting the requirements for miRNA annotation as outlined by Meyers *et al.*[18] These miRNAs were found to belong to seven miRNA families, with an additional two matches corresponding to as yet unnamed poplar miRNAs and a parologous miRNA with partial sequence homology to mtr-miR4414b. An additional 649 novel and distinct flax miRNA genes were identified to form from canonical hairpin structures and to have putative targets among the ~30,000 flax Unigenes. These novel flax miRNAs will need to be validated with high throughput sequencing of small RNA libraries.

Flax has fewer conserved microRNAs than expected

In comparison with other *in silico* analyses for miRNAs using either EST sequences, such as in Switchgrass with 44 families,[23] or genome survey sequences (GSS) such as in tobacco with 65 families,[48] the number of conserved miRNAs identified in flax appears low. However this is the first computational miRNA analysis performed on a complete genome. More comparable data has been observed in sugarcane where 14 families were identified using a combination of EST and GSS data, and from poplar, in which had 21 families identified from RNAseq data coupled to a computational approach, but only nine of these were conserved, the remainder being novel to poplar at that time.[49,50] The cassava genome which is of a similar size to that of flax had 17 miRNA families identified bioinformatically from ESTs.[51] The low number of conserved miRNAs identified from the genome analyses (two) may suggest that the computational program novoMIR is better suited towards EST analysis rather than that of genomes. It is also possible that during the rapid genomic changes in flax, which occur in response to stress, miRNA

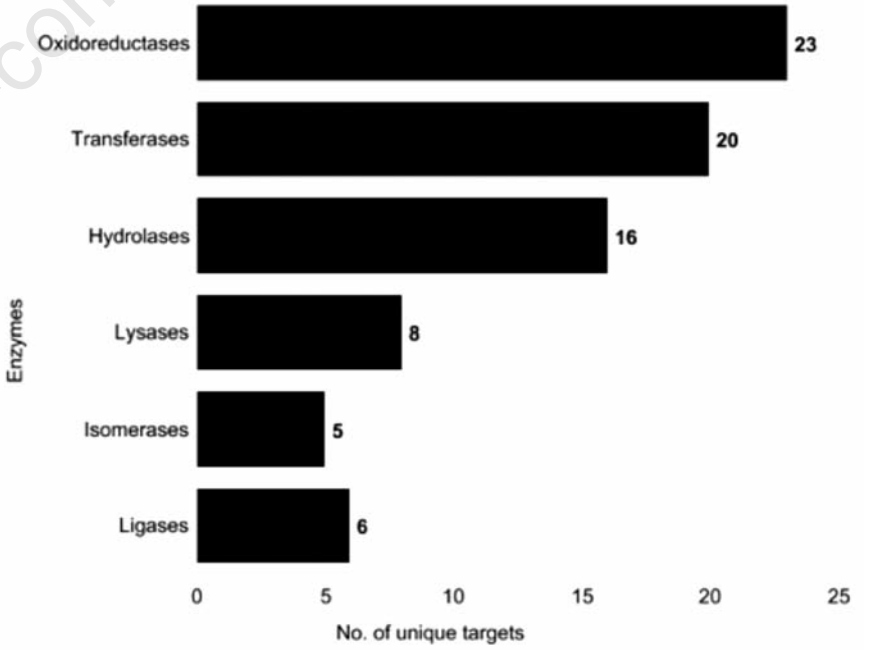

Figure 3. Enzyme Targets of flax miRNAs. Among the predicted targets, were 73 enzyme orthologs were annotated by Kyoto Encyclopedia of Genes and Genomes Automatic Annotation Server. Oxidoreductases and transferases were found to be the most prevalent.

sequences may co-evolve with their targets such that they become unidentifiable by means of sequence conservation analysis against known miRNAs in other plant species.[4,52] For that reason, we focused our analysis on miRNAs which were found to have putative targets among the flax Unigenes.

Advantages and disadvantages of source sequences

Computationally derived miRNAs and their targets are often confirmed experimentally to avoid false-positives.[18] This can be achieved using conventional cloning, Northern blots and cDNA cloning. However, in plants, these technologies are often not sensitive enough to detect miRNAs, as they are frequently underrepresented in the small RNA fraction.[47] In this study, the flax genome, Bethune v.09 and the flax seed EST database, UNIGENE were used to predict conserved and novel miRNAs. MiRNAs predicted from the flax genome which were not conserved were cross-referenced against the predicted genic regions. There are two benefits to this approach. First, it provides some confidence that the identified miRNAs are expressed, as their host genes are either expressed sequence tags or located within putatively transcribed regions, of which many have associated ESTs, and which also provides an alternative solution to the difficulties associated with large-scale experimental validation on expression of miRNAs. Secondly, expression data associated with these ESTs is readily available to assist in future miRNA experiments examining functional characteristics.[53]

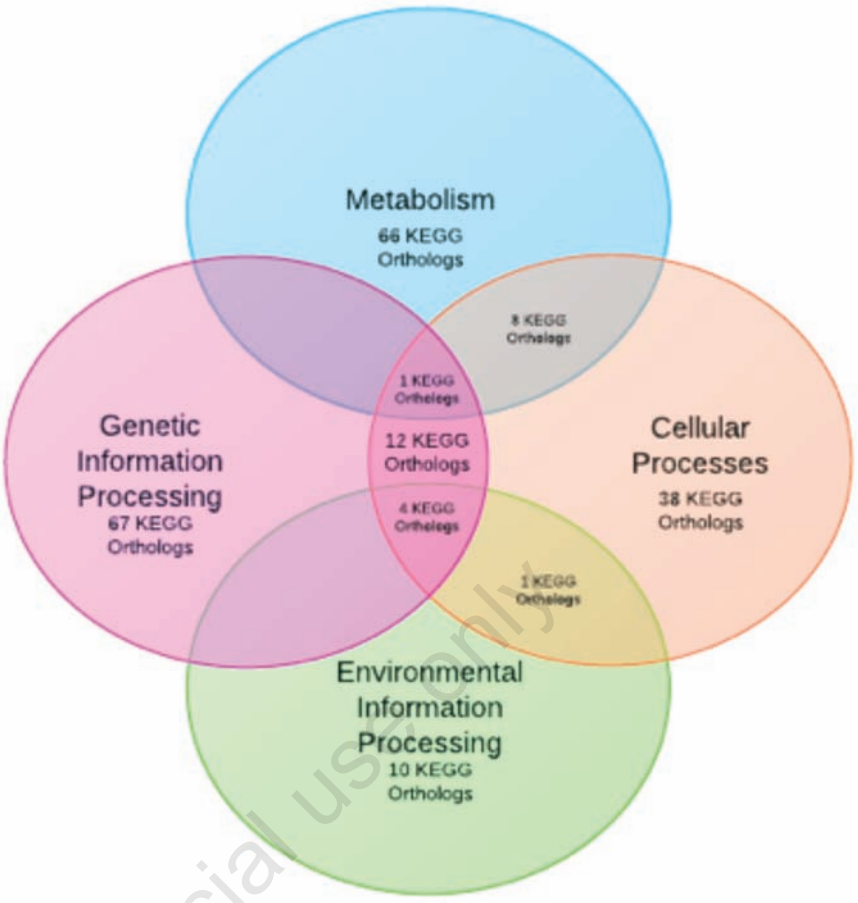

Figure 4. Flax microRNA Target Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Hierarchy The 175 predicted targets identified by KEGG Automatic Annotation Server to have orthologs with known functions were classified into the broad categories of Metabolism, Genetic Information Processing, Environmental Information Processing and Cellular Processes using KEGG BRITE.

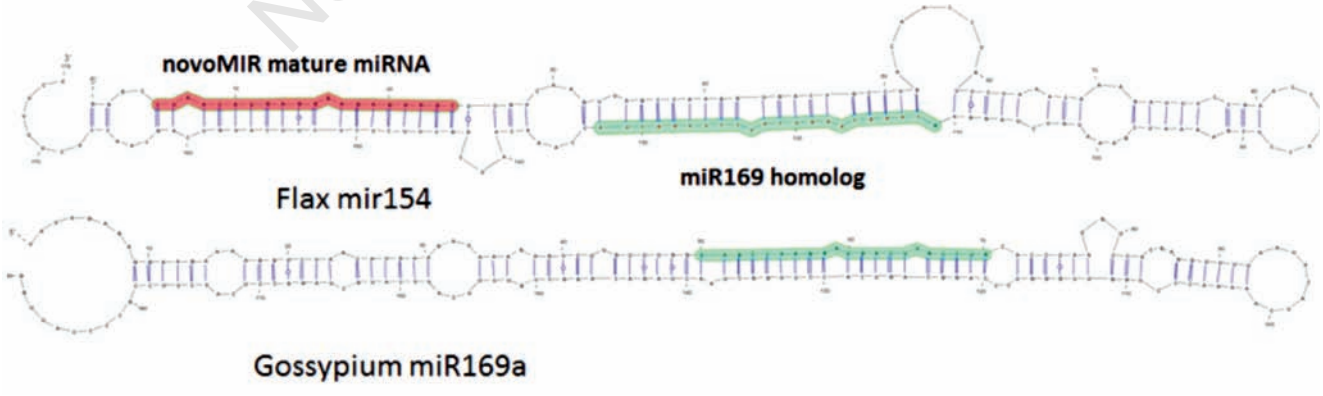

Figure 5. Conservation between the Linum mir169 (Supplementary File 1) predicted stem-loop precursor and Gossypium homolog. The novoMIR predicted mature miRNA sequence for flax mir154 has been highlighted as well as the mature miR169 homologous region. Although the miRNA169 could be found within the predicted hairpin, it was discarded by novoMIR due to the number of mismatches in the miRNA:miRNA* alignment. As expected sequence homology is seen in the paired sequences of the mature miRNA sequence. Although other parts of the sequence have evolved, the formation of the canonical stem-loop structure remains intact as well as the mature miRNA sequence.

Although computational analysis is often the most rationale first step in creating a miRNA profile, there are inherent limitations associated with this method. Some of these limitations are Phyla specific. In plants, it has been observed that while there can be high conservation in mature miRNA sequences, their precursor sequences are less likely to possess any sequence similarity to that of a related species, most likely due to the nature of adaptation in plants.[16] Consistent with previous findings, no pre-miRNAs were identified to have >85% identity to the pre-miRNA sequences belonging to the conserved mature miRNA sequences which have already been identified.[16] Therefore it was not surprising that of the many precursor sequences identified by NovoMIR only two pre-miRNAs were found to have any significant sequence similarity to that of another known pre-miRNAs, miR319 and miR169. As expected, the pre-miRNAs found to have any sequence similarity to those in flax were from a close relative of flax, poplar.

Flax miR319 associated with transposons

A single EST match with sequence similarity to a predicted gag-protease-integrase-RT-RNAseH and a reverse transcriptase was found among the many putative targets identified for miR319. As miR319 has been shown to be a highly conserved miRNAs responsible for targeting the LANCEOLATE (LA) class of TCP transcription factors, this association with another nucleic acid recognition protein suggests that a common motif may be responsible for the match observed in this data.[54] This is notable in light of recent finding that flax has relatively few known transposons and recent data suggesting a role for transposons in the evolution of miRNAs in plants.[26,55] Further analysis into the origin and function of miR319 in flax may provide additional insight into this phenomenon and that of the paucity of transposons in flax.

Flax miR169 associated with ribosome biogenesis

In miRBase v.18, miR169 represents one of the largest miRNA families, with over 200 different members (www.miRBase.org). The flax pre-miRNA 154 identified as belonging to this family was found to have limited sequence similarity to miRNA precursor ghb-miR169a (*Gossypium*) and ptc-miR169i while mature miRNA sequences from several plant species including Brassica (bna-miR169i), another oil seed crop, and Populus (ptc-169o), one of the closest relatives of flax with an annotated genome, had matches within the predicted flax miRNA precursor with less than two mismatches between nucleotides 111 and 132 of the precursor. Limited structural similarity was observed with the folded pre-miRNAs from flax and *Gossypium* (Figure 5).

The mature miRNAs from flax and Gossypium were identical and the placement of the miRNA sequence within these pre-miRNAs is shown in Figure 5 as well as the location of the novoMIR predicted mature miRNA sequence for flax mir154. This demonstrates the variability in the overall pre-miRNA sequence while maintaining the canonical hairpin structure with the conserved miRNA sequence not representing a large enough fraction of the pre-miRNA to be identified through homology searches alone. These data show the value in the *de novo* prediction method used here which will subsequently need to be supported by experimental data.

Conclusions

Although the numbers of conserved pre-miRNAs identified from the flax genome analysis are fewer than those seen in other genomes where a similar analysis has been performed, this may be due to significant differences in the composition of the flax transcriptome or suggest that the novoMIR algorithm is better suited to EST analysis. However, the number of total conserved miRNAs predicted from the EST database is similar to other miRNA analyses. High throughput sequencing of small RNAs in Bethune is needed to validate the novel miRNA sequences predicted here which do not have homologs among the current miRNA database on the basis of sequence conservation. The sequence, mapping and pRNAtarget data for all miRNAs derived from the flax genome and their precursors predicted by novoMIR are included in Supplementary File 14.

References

1. Durrant A. The environmental induction of heritable change in Linum. Heredity 1962; 17:27-61.
2. Chen Y, Lowenfeld R, Cullis CA. An environmentally induced adaptive (?) insertion event in flax. Int J Genet Mol Biol 2009;1:038-47.
3. Goldsbrough PB, Cullis CA. Characterisation of the genes for ribosomal RNA in flax. Nucleic Acids Res 1981;9:1301-10.
4. Cullis CA. Mechanisms and Control of rapid genomic changes in flax. Ann Bot 2005;95:201-6.
5. Schneeberger RG, Cullis CA. Specific DNA alterations associated with the environmental induction of heritable changes in flax. Genetics 1991;128:619.
6. Evans G, Durrant A, Rees H. Associated nuclear changes in the induction of flax genotrophs. Nature 1966;212:697-9.
7. Naqvi AR, Sarwat M, Hasan S, Choudhury NR. Biogenesis, functions and fate of plant microRNAs. J Cell Physiol 2012. [Epub ahead of print]
8. Zhang B, Pan X, Stellwag EJ. Identification of soybean microRNAs and their targets. Planta 2008;229:161-82.
9. Jagadeeswaran G, Saini A, Sunkar R. Biotic and abiotic stress down-regulate miR398 expression in Arabidopsis. Planta 2009;229:1009-14.
10. Cuperus JT, Fahlgren N, Carrington JC. Evolution and functional diversification of miRNA genes. Plant Cell 2011;23:431-42.
11. Khraiwesh B, Zhu JK, Zhu J. Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. Biochim Biophys Acta 2012;1819:137-48.
12. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281-97.
13. Reinhart BJ, Weinstein EG, Rhoades MW, et al. MicroRNAs in plants. Genes Dev 2002;16:1616-26.
14. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet 2009;10,:94-108.
15. Axtell MJ, Westholm JO, Lai EC. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. Genome Biol 2011;12:221.
16. Wang X, Zhang J, Li F, et al. MicroRNA identification based on sequence and structure alignment. Bioinformatics 2005;21:3610-4.
17. Ambros V, Bartel B, Bartel DP, et al. A uniform system for microRNA annotation. RNA 2003;9:277-79.
18. Meyers BC, Axtell MJ, Bartel B, et al. Criteria for annotation of plant microRNAs. Plant Cell 2008:20:3186-90.
19. Zhang B, Wang Q, Wang K, et al. Identification of cotton microRNAs and their targets. Gene 2007;397:26-37.
20. Jin W, Li N, Zhang B, et al. Identification and verification of microRNA in wheat (Triticum aestivum). J Plant Res 2008;121: 351-5.
21. Zhang W, Luo Y, Gong X, et al. Computational identification of 48 potato microRNAs and their targets. Comput Biol Chem 2009;33:84-93.
22. Gleave AP, et al. Identification and characterisation of primary microRNAs from apple (Malus domestica cv. Royal Gala) expressed sequence tags. Tree Genet Genomes 2007;4:343-58.
23. Xie F, Frazier TP, Zhang B. Identification and characterization of microRNAs and their targets in the bioenergy plant switch-

grass (Panicum virgatum). Planta 2010; 232:417-34.
24. Song C, Jia Q, Fang J, et al. Computational identification of citrus microRNAs and target analysis in citrus expressed sequence tags. Plant Biol 2010;12:927-34.
25. Teune JH, Steger G. NOVOMIR: De novo prediction of microRNA-coding regions in a single plant-genome. J Nucleic Acids 2010;1-11.
26. Venglat P, Xiang D, Qiu S, et al. Gene expression analysis of flax seed development. BMC Plant Biology 2011;11:74.
27. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011;39:D152-7.
28. Zhang Z, Yu J, Li D, et al. PMRD: plant microRNA database. Nucleic Acids Res 2010;38:D806-13.
29. Griffiths-Jones S, Saini HK, van Dongen S, Enright A.J. miRBase: tools for microRNA genomics. Nucleic Acids Res 2007;36: D154-8.
30. Gruber AR, Lorenz R, Bernhart SH, et al. The Vienna RNA Websuite. Nucleic Acids Res 2008;36:W70-4.
31. Giegerich R, Voss B, Rehmsmeier M. Abstract shapes of RNA. Nucleic Acids Res 2004;32:4843-51.
32. Steffen P, Voss B, Rehmsmeier M, et al. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics 2006;22:500-3.
33. Darty K, Denise A, Ponty Y. VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics 2009;25: 1974.
34. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;215:403-10.
35. Milev I, Yahubyan G, Minkov I, Baev V. miRTour: Plant miRNA and target prediction tool. Bioinformation 2011;6:248-9.
36. Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res 2011;39:W155-9.
37. Mao X, Cai T, Olyarchuk JG, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 2005;21:3787.
38. Moriya Y, Itoh M, Okuda S, et al. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 2007;35:W182-5.
39. Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res 2008;36:D480-4.
40. Nogueira FTS, Madi S, Chitwood DH, et al. Two small regulatory RNAs establish opposing fates of a developmental axis. Genes Dev 2007;21:750-5.
41. Mica E, Piccolo V, Delledonne M, et al. High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera. BMC Genomics 2009;10: 558.
42. Nogueira FT, Chitwood DH, Madi S, et al. Regulation of small RNA accumulation in the maize shoot apex. PLoS Genet 2009;5: e1000320.
43. Rhoades MW, Reinhart BJ, Lim LP, et al. Prediction of plant microRNA targets. Cell 2002;110:513-20.
44. Thakur V, Wanchana S, Xu M, et al. Characterization of statistical features for plant microRNA prediction. BMC Genomics 2011;12:108.
45. Devers EA, Branscheid A, May P, Krajinski F. Stars and symbiosis: MicroRNA- and microRNA*-mediated transcript cleavage involved in arbuscular mycorrhizal symbiosis. Plant Physiol 2011;156:1990-2010.
46. Brown JWS, Marshall DF, Echeverria M. Intronic noncoding RNAs and splicing. Trends in Plant Sci 2008;13:335-42.
47. Varkonyi-Gasic E, Hellens RP. Quantitative stem-loop RT-PCR for detection of microRNAs. Methods Mol Biol 2007;744: 145-57.
48. Frazier TP, Xie F, Freistaedter A, et al. Identification and characterization of microRNAs and their target genes in tobacco (Nicotiana tabacum). Planta 2010; 232:1289-308.
49. Li B, Yin W, Xia X. Identification of microRNAs and their targets from Populus euphratica. Biochem Biophys Res Commun 2009;388:272-7.
50. Zanca AS, Vicentini R, Ortiz-Morea FA, et al. Identification and expression analysis of microRNAs and targets in the biofuel crop sugarcane. BMC Plant Biol 2010;10:260.
51. Amiteye S, Corral JM, Sharbel TF. Overview of the potential of microRNAs and their target gene detection for cassava (Manihot esculenta) improvement. Afr J Biotechnol 2011;10:2562-73.
52. Voinnet, O. Origin, biogenesis, and activity of plant microRNAs. Cell 2009;136:669-87.
53. Roach MJ, Deyholos MK. Microarray analysis of developing flax hypocotyls identifies novel transcripts correlated with specific stages of phloem fibre differentiation. Ann Bot 2008;102:317-30.
54. Ori N, Cohen AR, Etzioni A, et al. Regulation of LANCEOLATE by miR319 is required for compound-leaf development in tomato. Nat Genet 2007;39:787-91.
55. Li Y, Li C, Xia J, Jin Y. Domestication of transposable elements into microRNA genes in plants. PloS One 2011;6:e19212.