Deposited research article

Conservation and divergence of microRNA families in plants

Tobias Dezulian[1], Javier F Palatnik[2], Daniel Huson[1] and Detlef Weigel[2]

Addresses: [1]Department of Algorithms in Bioinformatics, Center for Bioinformatics Tübingen, Tübingen University, D-72076 Tübingen, Germany. [2]Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076, Tübingen, Germany.

Correspondence: Tobias Dezulian. E-mail: dezulian@informatik.uni-tuebingen.de

Posted: 11 October 2005

Genome ***Biology*** 2005, **6:**P13

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/11/P13

Received: 10 October 2005

This is the first version of this article to be made available publicly.

Research

Conservation and divergence of microRNA families in plants

Tobias Dezulian[1], Javier F Palatnik[2], Daniel Huson[1] and Detlef Weigel[2]

[1]Department of Algorithms in Bioinformatics, Center for Bioinformatics Tübingen, Tübingen University, D-72076 Tübingen, Germany.

[2]Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany.

Correspondence: Tobias Dezulian. E-mail: dezulian@informatik.uni-tuebingen.de

Email addresses for all authors:

TD: dezulian@informatik.uni-tuebingen.de

JFP: javier.palatnik@tuebingen.mpg.de

DH: huson@informatik.uni-tuebingen.de

DW: weigel@tuebingen.mpg.de

Abstract

Background:

MicroRNAs (miRNAs) are 20 to 24 nucleotides short RNAs involved in posttranscriptional regulation in plants and animals. MiRNAs are processed from larger precursors with extensive secondary structure. In plants, a total of 286 miRNA genes in Arabidopsis, rice and maize had been identified by March 2005, clustered in 43 families.

Results:

Here, we report the bioinformatic identification of 200 members of the 43 miRNA families in the genomes of maize, sorghum, medick and poplar. Furthermore, we report evidence for expression of 37 miRNA precursors that are present in EST collections of soybean and sugarcane. We have used the enlarged data set to systematically analyze several parameters of the plant precursors including stem length, conservation of the precursors and variation in the secondary structure of the miRNA along the precursor.

Conclusion:

Based on this 83% increase in available miRNA precursor sequences, we present an improved view of phylogenetic distribution, positional nucleotide preference, structural features and conservation of miRNA genes. Our results suggest that there are two different classes of plant miRNA precursors. The most abundant class includes precursors that have only two strongly conserved regions, corresponding to the mature miRNA and its complementary sequence. A less frequent class, which includes the miRNA families miR159/319 and miR394, displays two additional conserved sequence blocks. These precursors have larger stems with more extensive secondary structure.

Background

MicroRNAs (miRNAs) are products of endogenous non-coding RNA genes that perform recently appreciated regulatory roles in plants and animals. MiRNAs are typically 20 to 24 nucleotides in length and exhibit different degrees of sequence complementarity to target RNAs, which they regulate at the posttranscriptional level [1]. Effectively, miRNAs confer guidance specificity to an effector complex that negatively regulates expression of these target mRNAs. In animals, miRNAs usually recognize several target sequences in the 3' UTR and cause translation inhibition, while plant miRNAs often recognize one target site in the coding region and guide the mRNA to cleavage. It is thought that the higher degree of complementarity between plant miRNAs and their targets favors the latter mechanism.

MiRNA biogenesis starts with the transcription of a large primary transcript, which is typically RNA polymerase II dependent and may be spliced. The mature pri-miRNA transcripts is capped and polyadenylated. Within this transcript the pre-miRNA is found as part of a fold-back structure in which the miRNA is located as part of the stem. The region of the precursor that pairs to the miRNA is called the miRNA*. While primary transcripts in animals can contain several different miRNA precursors, plant transcripts usually contain a single pre-miRNA. The function, if any, of these larger transcripts is currently unknown. MiRNA precursors are processed by ribonuclease III enzymes (Drosha and Dicer in animals and DICERLIKE1 in Arabidopsis) to release the mature miRNAs. MiRNAs are subsequently incorporated into an RNA induced silencing complex (RISC) on which they confer sequence specificity to the scanning of target mRNAs.

More than two hundred human miRNAs and one hundred Arabidopsis sequences are annotated in the current edition of the miRNA registry (http://microrna.sanger.ac.uk/registry/). Plant miRNAs have been identified using one of three primary strategies. The first relies on the direct cloning of small RNAs. Several labs have prepared small RNA libraries from Arabidopsis and one from rice including different tissues and conditions [2-4]. A second approach is based on the extensive conservation of the miRNAs during evolution. 20mers that are conserved between Arabidopsis and rice with surrounding sequences that are able to form fold-back structures have allowed the identification of several new miRNAs and the postulation of many others [5, 6]. A third approach has been the identification of

miRNAs through forward genetics, an approach that led to the first identification of small RNAs in animals [7]. Although there have been a few cases thus far where a miRNA family is unique to Arabidopsis or rice, the majority of validated plant miRNAs are largely conserved across the plant kingdom.

Plant miRNAs are usually encoded by small gene families of up to 14 members [8]. A miRNA family usually includes several mature miRNAs of identical or nearly identical sequence. In general, at least one member of each family has been validated experimentally, while the others have often been obtained as bioinformatic predictions based on sequence similarity of the miRNA and the ability of the surrounding region to adopt a fold-back structure. In most of the cases, there is some heterogeneity in the ends of miRNAs belonging to the same family. It is still not known whether these differences reflect functional specialization. The fold back structures, which are the key structural feature of the miRNA precursors, are thought to contain spatial cues that direct the appropriate miRNA processing. Precursors are quite uniform in length among animal miRNAs (around 70 nucleotides), but are variable in plants (ranging from 50 to 500 nucleotides). Interestingly, engineering the miRNA sequence of the Arabidopsis miR171 has successfully led to an artificial miRNA that can guide cleavage of an arbitrarily chosen foreign target, suggesting that the specific determinants of the position of the miRNA sequence along the precursor are located only in the fold-back structure and not in the sequence of the miRNA itself [8]. Two recent reports have shown different evolutionary patterns for miRNA precursors in plants. Reinhardt et al. [3] reported for many families that only the mature miRNA and the miRNA* was significantly conserved between Arabidopsis and rice for several miRNAs. In contrast, Palatnik et al. [7] found extensive conservation of the miR319 precursor between dicots and monocots.

As most of the information regarding plant miRNAs has been based on the two model species Arabidopsis and rice, we decided to extend the systematic identification of miRNAs to other plant species. Therefore we have performed a thorough search based simultaneously on sequence similarity and precursor structure. To date, a total of 286 plant miRNAs in the dicot Arabidopsis and the monocots rice and maize have been identified and deposited in the miRNA registry [9] responsible for name assignment. These miRNAs cluster into 43 families, based on sequence identity (or very close similarity) of the mature miRNA. At least one representative of each family has been

experimentally verified - whereas other family members have often been identified by computational approaches. Here, we report on the identification of 237 additional miRNAs, all homologous to members of published families, based on a stringent sequence similarity search in combination with a structural filter for hairpin structures and manual inspection. We provide an overview of characteristics referring to the conservation and divergence of miRNAs in plants as provided by the now enlarged set of miRNA sequences.

Results and Discussion

Computational identification of miRNA homologs

In order to identify miRNA genes from different species we used the set of 286 plant miRNAs (112 from Arabidopsis, 134 from rice and 40 from maize) available from the miRNA registry [9] Release 5.1 (March 2005) as query set. We used NCBI BLAST [10] with an E-value cutoff of 10 for a similarity search against the available genomic sequences of sorghum, maize, medick and poplar, as well as all NCBI EST databases. Of the four species used for genomic DNA searches, only poplar has been completely sequenced. It has been estimated that about two thirds of the genic fraction of the sorghum and maize genomes has been obtained [11, 12]. All BLAST hits were folded using RNAfold from the Vienna RNA package [13] and checked for adoption of a stem-loop-stem structure. We also applied a second structural filter and required that the newly identified sequences are similar in the miRNA-miRNA* region to the one described in RFAM. To this end, we required that at least the same number of interacting bases between the miRNA and miRNA* are found in the different miRNAs. All miRNA candidates that passed this test were manually inspected for homology with the query.

Using this strategy we were able to identify 237 new miRNA genes, 200 of them from the available genome sequences and 37 from expressed sequences (Figure 1a, Supplementary table S1) - thus increasing the number of available miRNA genes by roughly 83%. The strategy used to generate a set of miRNA homolog candidates is mainly based on sequence similarity by BLAST, with an E-value

cutoff of 10. In order to validate this approach we wanted to compare the similarity inside one family and across different families. We started our search with all precursors deposited in RFAM. There is some heterogeneity in the sequences outside the miRNA-miRNA* for different precursors. In some cases the sequences deposited with the miRNA registry end immediately beyond the miRNA sequence; in other cases they continue for an arbitrary number of nucleotides, since the actual length of the transcribed miRNA precursor is in many cases not known. To standardize our parameters, we decided to only take into account the miRNA/loop/miRNA* portion of each miRNA precursor and to normalize the scores when conducting pairwise blast comparisons to avoid any bias introduced by differential family size. As depicted in Figure 1b, the pairwise BLAST scores vary across families, but we obtained at least a pairwise comparison score of 25 points for any pair of miRNAs within a family. As a control we shuffled the nucleotides of each precursor and did not get any significant scores (data not shown).

Arabidopsis microRNAs for which we did not detect homologs in other species by our approach likely reflect two scenarios. One possibility is that these miRNAs have recently appeared in Arabidopsis, as was shown for miR161 and miR163 [14]. A second alternative is that the original Arabidopsis sequences do not constitute bona fide miRNAs, since DICER-LIKE1 dependent biogenesis has not been determined for all small RNAs in the miRNA registry.

The number of miRNA family members varies greatly across families. In some cases, like miR162, there are only two members in Arabidopsis, while there are 14 members in the miR169 family. Thus, we decided to ask whether the number of family members also varies across species or between monocot and dicot clades. We plotted the number of miRNA members in each family found in the sequenced (or almost fully sequenced) genomes from dicots and monocots. From dicots we used Arabidopsis and poplar, and from monocots we used rice, sorghum and maize. Surprisingly, we found that the number of miRNA genes is conserved in each family of miRNAs. This is interesting because the genome size of monocot species is often much larger than for dicots. Importantly, none of the families analyzed behaves as an outlier (Fig. 1c and Fig. 2).

Conservation and divergence of miRNA precursors

MicroRNAs in plants and animals have distinct structural features and differ, for example, in the length of their precursors. The fold-back of a miRNA precursor in animals is usually around 70 nucleotides in length, while in plants it can be anywhere between 50 and 500 nucleotides [1]. In a first step to characterize the plant precursors, we looked at the length distribution across miRNA families. We found that several miRNA families such as miR164 or miR408 have variable fold-back sizes (Fig. 2b). Other miRNAs families, such as miR319, showed little variation (Fig. 2b). These results suggest that distinct plant miRNAs may be subject to different structural constraints.

Next, we analyzed in more detail the structural conservation of plant miRNA precursors. We compared the sequences from different miRNAs belonging to the same family across different species. For this analysis we used only those *enlargeable* miRNA precursors for which the segment 50 nucleotides upstream and downstream of the miRNA and miRNA* was available. To compare the miRNA precursors we used the multiple alignment program T-Coffee [15] for alignment of this segment and for visualization of conservation at each position. As expected, the miRNA sequence itself shows extreme evolutionary conservation in all families. We found maximally 2 nucleotide deviations from the consensus sequence for a miRNA family. The miRNA* sequence is second best in conservation, being constrained by the necessity to form a double stranded RNA intermediary together with the miRNA sequence. Compensatory mutations due to wobble pairing and mutations at bulge positions are tolerated.

The very interesting result was derived from an analysis of the conservation of the miRNA precursor segment located between the miRNA and miRNA*, termed the loop. In most cases, such as miR160 and miR164, we found that there was essentially no conservation of this sequence between different species (Fig. 3a and 3b). Astonishing exceptions to this divergence pattern are the families miR159, miR319 and miR394. In these families we found that the loop region of the fold back displays an unusual amount of conservation. Especially in the families miR159 and miR319 there were two distinct

other conserved blocks that gave signals similar to that of the miRNA and miRNA* (Fig. 3c and 3d). In the case of miR159 and miR319, this interesting, shared pattern of conservation might not be accidental, since the mature miR159 and miR319 sequences are also closely related, even though they appear to have largely non-overlapping targets [7, 16, 17]. Next, we decided to use the precursor sequence without the miRNA/miRNA* to avoid any bias due to these sequences and we performed a BLAST search against our miRNA precursor database. When we used the sequence of the miR319 precursor as our query, we found that the best pairwise hit is a miR159 and vice versa. An alignment of all enlargeable miR159 sequences together with all enlargeable miR319 sequences yielded the same conservation pattern with four distinct conserved blocks (Fig. 3e). Taking into account the conservation of the miRNA sequence and the extraordinary conservation pattern in the fold-back precursor, we conclude that these miRNAs most likely have a common ancestor.

Why do the miR159/319 and 394 families have additional blocks of conserved sequence in their precursors? One possibility is that these segments code for a second miRNA, but this seems not very likely, as the sequence conservation is lower than found for authentic miRNA/miRNA* pairs. In addition, a search for putative target mRNAs using previously established rules [18] yielded no promising candidates. Furthermore, the few hits obtained in other species were not in orthologous genes. A more likely possibility may be that these segments are required for adequate miRNA biogenesis and therefore have a structural function.

Sequence signals

Some primary sequence bias in miRNAs has been detected, with U being the most common base at the extreme 5' end. Having this now-enlarged set of miRNA precursors at hand, we constructed position-weight-matrices (PWM) for all mature miRNA (miRNA*) sequences of length 21 for which 50 nucleotides upstream and downstream of the mature miRNA (miRNA*), respectively, were available [19]. We first analyzed the 307 miRNA enlargeable precursors (out of the 523 overall) with a 21 nucleotide mature sequence and graphed nucleotide bias per position (Fig. 4a), confirming the clear U preference at the first position. To exclude the possibility that this result is influenced by different miRNA family sizes, we repeated this exercise after having selected only one enlargeable representative per family at random and constructing a PWM for this set of 35 representatives (Figure

4c). Note that the G and C preferences at positions 8 and 19, respectively, visible in Figure 4a are mostly due to the bias introduced by the large families miR169 and miR166, and they disappear after using only a single representative per family (Fig. 4a, 4c). We found weak sequence signatures that have not been reported before, such as a pyrimidine preference at the first position downstream of the mature miRNA and a thymine preference at the fifth position downstream of the mature miRNA, although these signals were much weaker than the previously described U at position 1. Figure 4b and 4d show analogous PWMs for enlargeable miRNA* sequences of length 21, using all 187 available sequences and the 28 sequences derived by selecting one representative per family at random, respectively.

We next analyzed the secondary structure of the miRNA and miRNA*. We scored the strength of the bond at each position of the miRNA and the miRNA*, scoring each from its 5' end, respectively, to compensate for variation in miRNA length and the effect of asymmetric bulges. We gave different values to the possible base pairs: GC a score of 3; AU, 2; GU, 1; unpaired nucleotides scored 0 at that position. We found that the 5' nucleotide of the miRNA scores an average strength of roughly 1.6, while the 5' nucleotide of the miRNA* scores on average a bond strength of about 2.4. This result indicates that the first nucleotide of the miRNA is more likely to be unpaired than the first nucleotide of the miRNA*. This finding is consistent with previous reports in animals which claimed that the protein complex in charge of loading RISC can discriminate the miRNA from miRNA* on the basis of the different stability of the 5' end. This strand selection theory is consistent with our results in plants.

Conclusions

Employing a similarity search of genomic and EST sequences with subsequent structural verification, we have been able to increase the number of plant miRNAs by 83% to 523 miRNA genes. Our analysis of this enlarged set has led to the following conclusions:

1. In contrast to animals, plant miRNA precursors were already known to be more variable in length. While we can confirm that there is size variation both across and between families, we also found that not all families are equally variable. In some families, all members are uniform in size. This phenomenon might reflect evolutionary trajectories and/or differential functional constraints.

2. The distribution of miRNA families is similar between monocots and dicots.

3. There is no obvious sequence bias within the primary sequence of different miRNAs, with the exception of the already known U in the first position. Consistent with the strand selection model for miRNA incorporation into RISC, we observe a bias in bond strength between the 5' end of the miRNA and 5' end of the miRNA*.

4. It seems that there are two classes of precursors with different structural properties. The most abundant class includes precursors that have only two strongly conserved regions or blocks comprising the miRNA and miRNA*. The foldbacks of these precursors contain a short stem consisting mainly of the miRNA/miRNA duplex. A second and less frequent class, which includes the miRNA families miR159/319 and miR394, display four conserved sequence blocks. This is reflected in the secondary structure of these precursors, which typically contain two adjacent, strongly paired stem segments. This likely reflects a processing mechanism that requires two consecutive steps by DICER-LIKE enzymes, in a similar way to the progressive action of DICER in siRNA production.

Materials and methods

For the genomic homology search we used the following sequences: the poplar (Populus trichocarpa) genome was downloaded from the DoE Joint Genome Institute and Poplar Genome Consortium web page at http://genome.jgi-psf.org/Poptr1/Poptr1.download.html on 10th of March 2005 (Version 1.0 preliminary draft). The maize (Zea mays) genome was downloaded from the MAGI website of the Iowa State University on 10th of March 2005 (MAGI Version 3.1 Contigs w/ Non-repetitive Singletons) [20]. The sorghum (Sorghum bicolor) genome was downloaded from the MAGI website of the Iowa State University on 10th of March 2005 (SAMI Version 2.0 Contigs w/ Singletons) [11]. The genomic medick sequences were downloaded on 10th of March 2005 from the NCBI Genome Survey Sequence database at ftp://ftp.ncbi.nih.gov/blast/db/ (Files "gss.0X.tar.gz"). For the homology search in EST sequences we used the NCBI EST database [21], downloaded on 10th of March 2005.

The 286 available plant miRNA sequences used as queries for our search were downloaded on 10th of March 2005 from the miRNA registry [9] at http://microrna.sanger.ac.uk/sequences/index.shtml (Release 5.1). All miRNA homologs identified in this approach have been deposited with the miRNA registry (Suppl. table S1).

For the pairwise BLAST comparisons we normalized the scores by setting a virtual (effective) database size equal to the length of the current NCBI NR/NT database (13,371,533,914 nucleotides). We find this useful since the BLAST E-value and score are directly related to the effective database length and omitting this would introduce a bias in similarity due to differential family size.

Acknowledgements

We thank Patrick S. Schnable and Yan Fu for information regarding the Maize and Sorghum genomes. We especially thank Rebecca Schwab and Stefanie Röhm for valuable discussions and suggestions.

References

1. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.
2. Llave C, Kasschau KD, Rector MA, Carrington JC: **Endogenous and silencing-associated small RNAs in plants**. *Plant Cell* 2002, **14**(7):1605-1619.
3. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP: **MicroRNAs in plants**. *Genes Dev* 2002, **16**(13):1616-1626.
4. Sunkar R, Zhu JK: **Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis**. *Plant Cell* 2004, **16**(8):2001-2019.
5. Jones-Rhoades MW, Bartel DP: **Computational identification of plant microRNAs and their targets, including a stress-induced miRNA**. *Mol Cell* 2004, **14**(6):787-799.
6. Wang XJ, Reyes JL, Chua NH, Gaasterland T: **Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets**. *Genome Biol* 2004, **5**(9):R65.
7. Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, Carrington JC, Weigel D: **Control of leaf morphogenesis by microRNAs**. *Nature* 2003, **425**(6955):257-263.
8. Parizotto EA, Dunoyer P, Rahm N, Himber C, Voinnet O: **In vivo investigation of the transcription, processing, endonucleolytic activity, and functional relevance of the spatial distribution of a plant miRNA**. *Genes Dev* 2004, **18**(18):2237-2242.
9. Griffiths-Jones S: **The microRNA Registry**. *Nucleic Acids Res* 2004, **32**(Database issue):D109-111.
10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
11. Bedell JA, Budiman MA, Nunberg A, Citek RW, Robbins D, Jones J, Flick E, Rholfing T, Fries J, Bradford K *et al*: **Sorghum genome sequencing by methylation filtration**. *PLoS Biol* 2005, **3**(1):e13.
12. Fu Y, Emrich SJ, Guo L, Wen TJ, Ashlock DA, Aluru S, Schnable PS: **Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes**. *Proc Natl Acad Sci U S A* 2005.
13. Hofacker IL, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster: **Fast Folding and Comparison of RNA Secondary Structures**. *Monatshefte f Chemie* 1994, **125**:167-188.
14. Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC: **Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana**. *Nat Genet* 2004, **36**(12):1282-1290.
15. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**(1):205-217.
16. Achard P, Herr A, Baulcombe DC, Harberd NP: **Modulation of floral development by a gibberellin-regulated microRNA**. *Development* 2004, **131**(14):3357-3365.
17. Millar AA, Gubler F: **The Arabidopsis *GAMYB*-like genes, *MYB33* and *MYB65*, are microRNA-regulated genes that redundantly facilitate anther development**. *Plant Cell* 2005, **17**(3):705-721.
18. Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, Weigel D: **Specific effects of microRNAs on the plant transcriptome**. *Dev Cell* 2005, **8**(4):517-527.
19. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, **14**(6):1188-1190.
20. Emrich SJ, Aluru S, Fu Y, Wen TJ, Narayanan M, Guo L, Ashlock DA, Schnable PS: **A strategy for assembling the maize (Zea mays L.) genome**. *Bioinformatics* 2004, **20**(2):140-147.
21. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags"**. *Nat Genet* 1993, **4**(4):332-333.

Figure captions

Figure 1. Overview of microRNA families.

(a) Number of microRNA genes per family. Blue indicates genes described in the RFAM microRNA registry (Release 5.1) and used as starting query, red indicates newly identified genes in different genome sequences and yellow newly identified genes obtained from expression sequence databases.
(b) Pairwise BLAST score of each stem-loop precursor sequence (miRNA + loop + miRNA*) compared to the different members of the same family. Each data point is plotted in black. Statistical symbols are drawn in gray: maximal and minimal values by horizontal marks, first and 99th percentile by crosses, mean value by a square. A gray box covers the range from 25% to 75% with whiskers extending to 10% and 90%.
(c) Histogram of microRNA genes (dicots: greenish/blue, monocots: reddish/yellow)

Figure 2. Conservation.

(a) Number of monocot family members (rice, maize and sorghum) plotted against number of dicot family members (Arabidopsis, poplar).
(b) Stem-loop length variation for different microRNA families. Statistical symbols as in Fig. 1b.

Figure 3. Alignments of selected microRNA families.

Alignments of microRNA genes are displayed. The alignment software used, T-Coffee, provides an algorithm and a coloring scheme to indicate degree of conservation: red/yellow/green/blue color symbolizes excellent/good/average/bad conservation, respectively. The mature microRNA is marked by a red rectangle and the sequence segment pairing to the microRNA is marked by a blue rectangle. The alignments have not been curated manually.
(a) Alignment of members of the miR160 family.
(b) Alignment of members of the miR164 family.
(c) Alignment of members of the miR319 family.
(d) Alignment of members of the miR159 family.
(e) Alignment of members of the miR159 together with members of the miR319 family.

Figure 4. Sequence logos.

Position weight matrices (PWMs) of miRNA genes are displayed using the software WebLogo. The y-axis indicates the total number of informative bits [19] for each position. Within each column, the fraction of height covered by each nucleotide is equal to its fraction of occurrences at the corresponding position.

(a) and (c) display sequence logos of the mature miRNA plus 50 nucleotides 5' and 3' across all microRNA families and across one randomly chosen representative for each family, respectively. The mature miRNA is underlined with a red bar. Only miRNA precursors that could be extended by 50 nucleotides to 5' and 3' of the mature miRNA and yield a mature miRNA of length 21 have been used for this analysis.

(b) and (d) display sequence logos of the miRNA* plus 50 nucleotides 5' and 3' across all miRNA families and across one randomly chosen representative for each family, respectively. The miRNA* is underlined with a blue bar. Only miRNA precursors that could be extended by 50 nucleotides to 5' and 3' of the miRNA* and yield a miRNA* of length 21 have been used for this analysis.

Figure 5. Bond strength of miRNA and miRNA*.

The average score (as described in the text) for the hydrogen bonds at the first 10 positions of the miRNA and the miRNA*, respectively, is displayed in blue and green. In red, the difference between the score for the miRNA and the miRNA* is plotted.

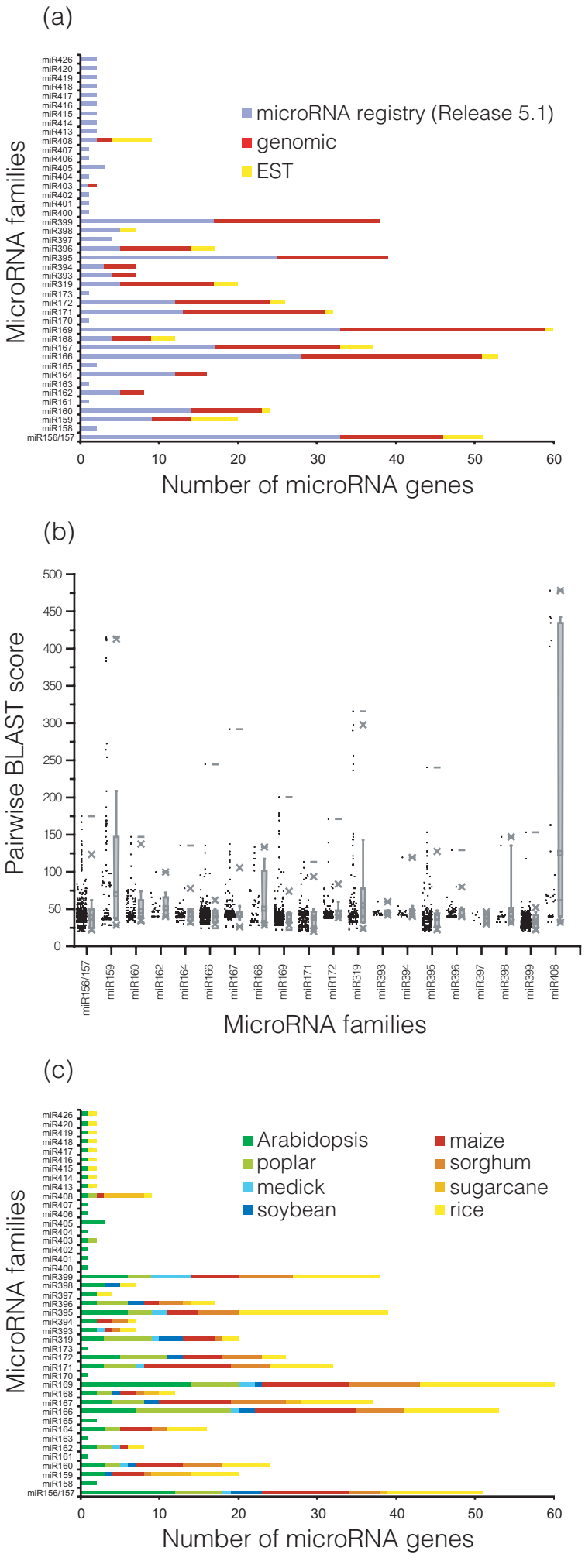
(a)
MicroRNA families
microRNA registry (Release 5.1)
genomic
EST
Number of microRNA genes
(b)
Pairwise BLAST score
MicroRNA families
(c)
MicroRNA families
Arabidopsis
poplar
medick
soybean
maize
sorghum
sugarcane
rice
Number of microRNA genes

Figure 1

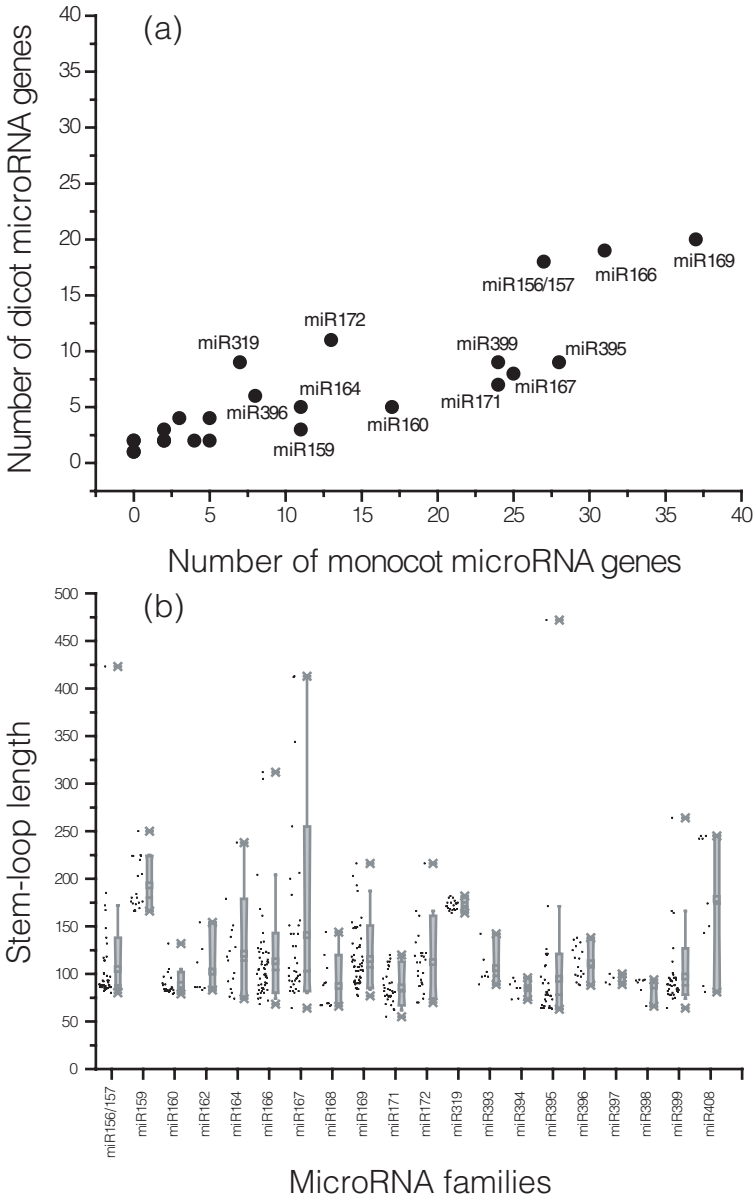
(a)
Number of dicot microRNA genes
Number of monocot microRNA genes
miR156/157
miR166
miR169
miR172
miR319
miR399
miR395
miR164
miR167
miR396
miR171
miR160
miR159
(b)
Stem-loop length
MicroRNA families
miR156/157
miR159
miR160
miR162
miR164
miR166
miR167
miR168
miR169
miR171
miR172
miR319
miR393
miR394
miR395
miR396
miR397
miR398
miR399
miR408

Figure 2

(a)

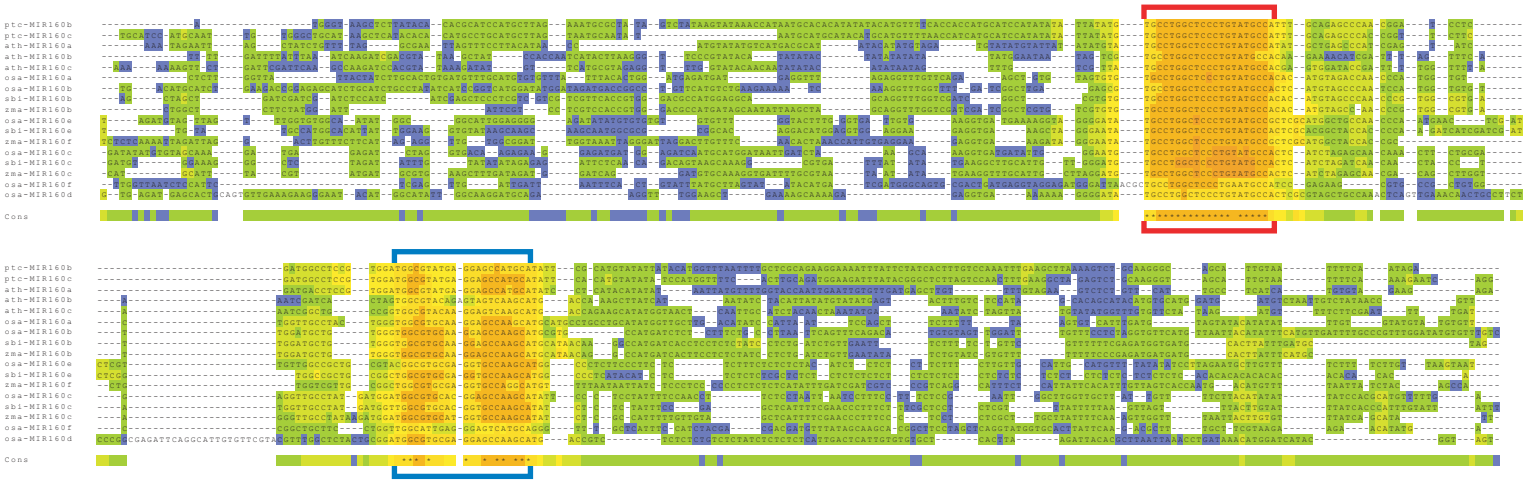

Figure 3

(b)

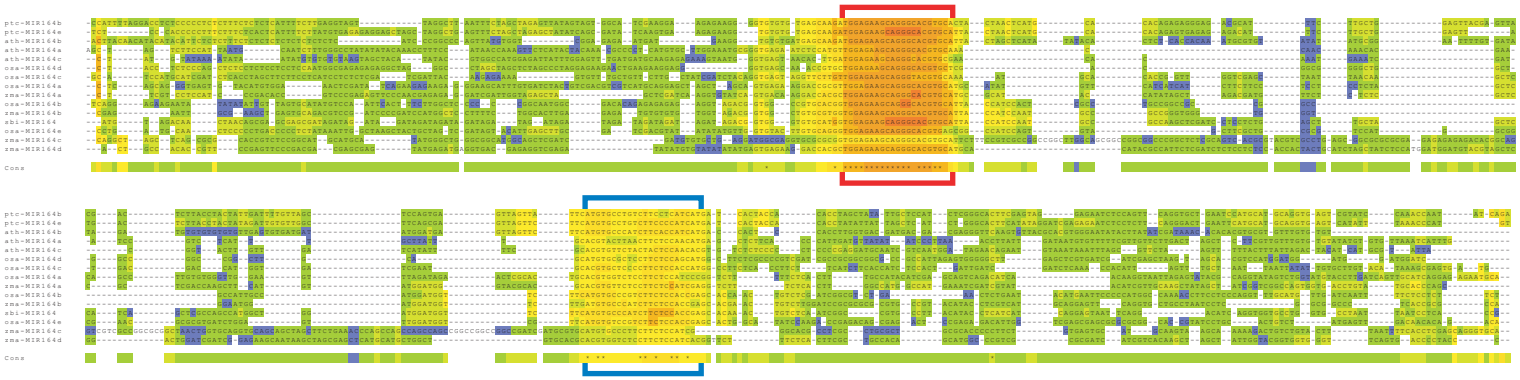
ptc-MIR164b
ptc-MIR164e
ath-MIR164b
ath-MIR164a
ath-MIR164c
osa-MIR164d
osa-MIR164c
osa-MIR164a
zma-MIR164a
osa-MIR164b
zma-MIR164b
sbi-MIR164
osa-MIR164e
zma-MIR164c
zma-MIR164d
Cons

Figure 4

(c)

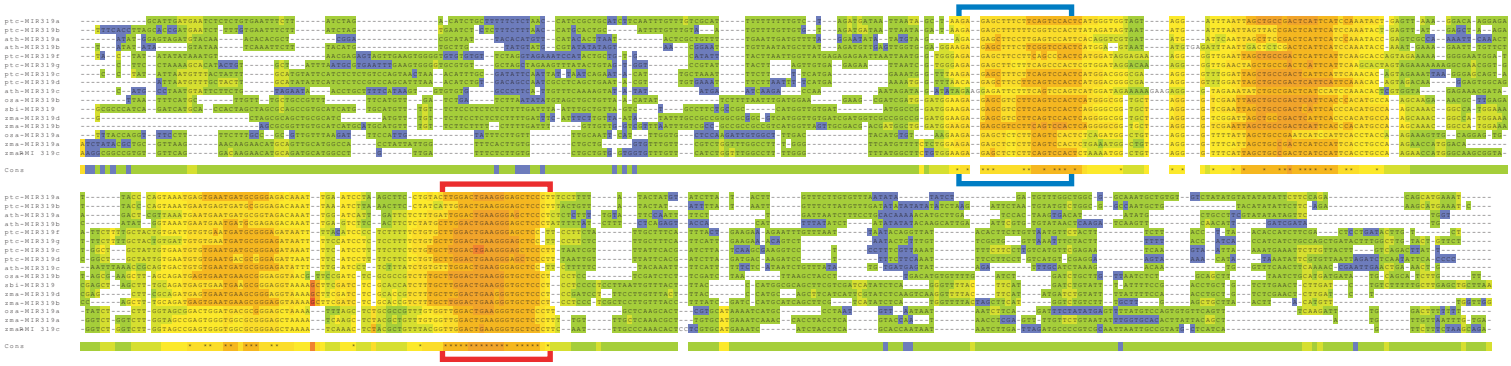

Figure 5

(d)

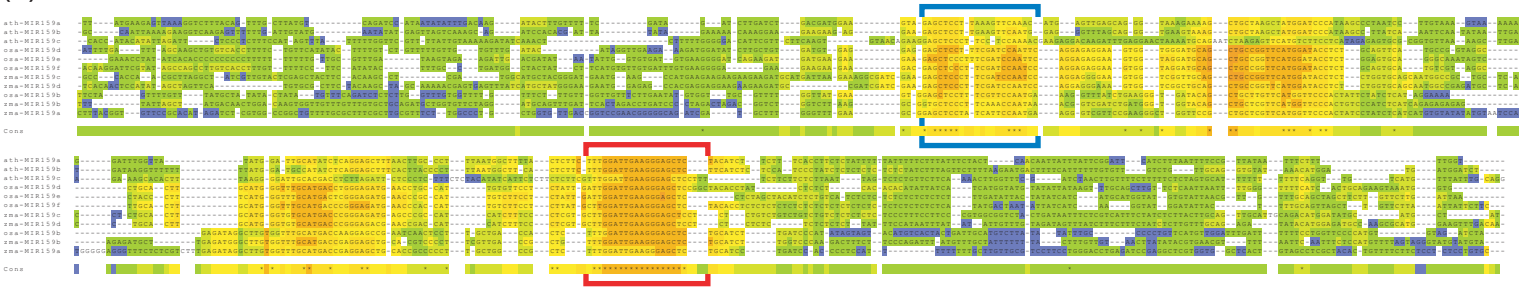

Figure 6

(e)

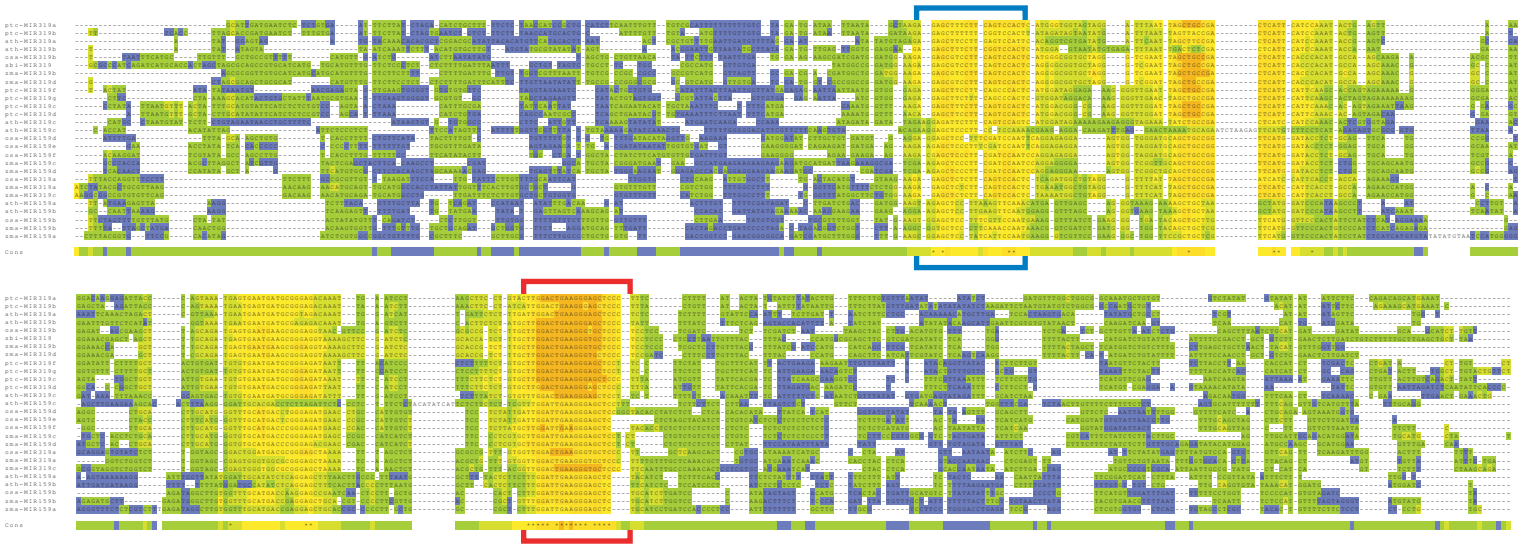

Figure 7

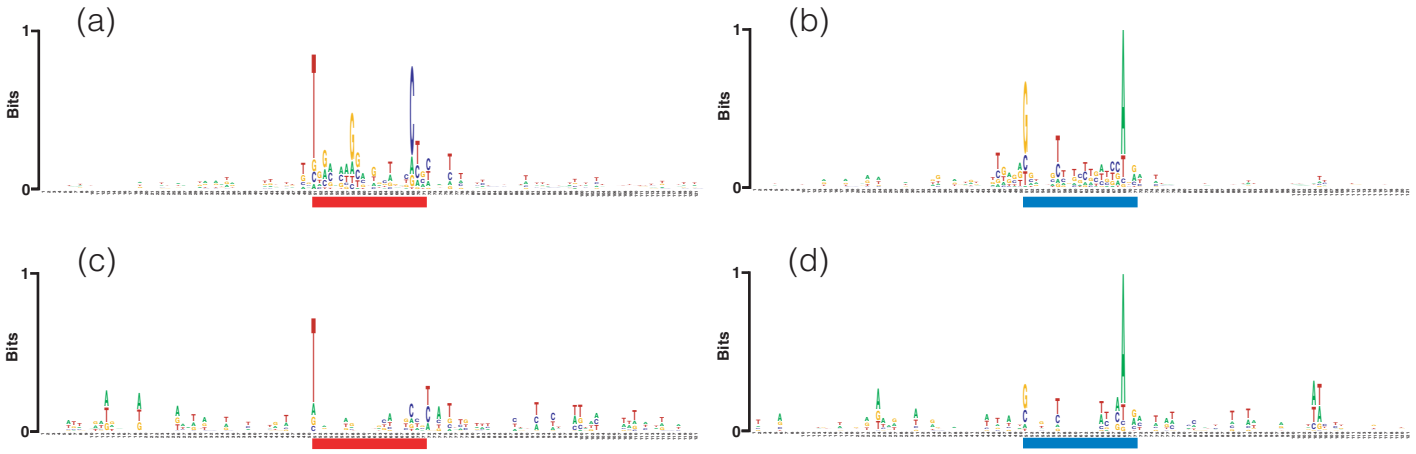

Figure 8

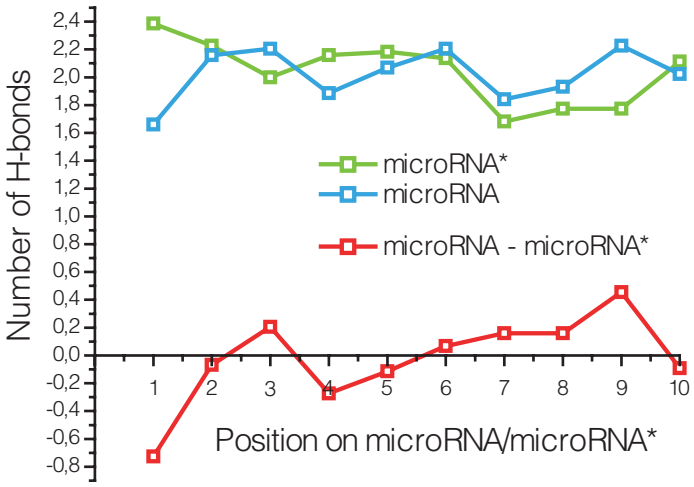
Number of H-bonds
2,4
2,2
2,0
1,8
1,6
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
1
2
3
4
5
6
7
8
9
10
microRNA*
microRNA
microRNA - microRNA*
Position on microRNA/microRNA*

Figure 9

Additional files provided with this submission:

Additional file 2 : PopulusTrichocarpa_homologs.fa : 15Kb
http://genomebiology.com/imedia/6645066037821926/sup2.FA

Additional file 1 : Supp_Table_S1_identified_miRNA_homologs.pdf : 131Kb
http://genomebiology.com/imedia/1074501341782177/sup1.PDF