BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 9 2011, pages 1215–1223
doi:10.1093/bioinformatics/btr132

Genome analysis

Advance Access publication March 17, 2011

SplamiR—prediction of spliced miRNAs in plants

Christoph J. Thieme[†,‡], Lydia Gramzow[†], Dajana Lobbes and Günter Theißen[*]

Department of Genetics, Friedrich Schiller University Jena, Philosophenweg 12, 07743 Jena, Germany

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MicroRNAs (miRNAs) are important regulators of biological processes in plants and animals. Recently, miRNA genes have been discovered, whose primary transcripts are spliced and which cannot be predicted directly from genomic sequence. Hence, more sophisticated programs for the detection of spliced miRNAs are required.

Results: Here, we present the first method for the prediction of spliced miRNAs in plants. For a given genomic sequence, SplamiR creates a database of complementary sequence pairs, which might encode for RNAs folding into stem–loop structures. Next, *in silico* splice variants of database sequences with complementarity to an mRNA of interest are classified as to whether they could represent miRNAs targeting this mRNA. Our method identifies all known cases of spliced miRNAs in rice, and a previously undiscovered miRNA in maize which is supported by an expressed sequence tag (EST). SplamiR permits identification of spliced miRNAs for a given target mRNA in many plant genomes.

Availability: The program is freely available at http://www.uni-jena.de/SplamiR.html.

Contact: guenter.theissen@uni-jena.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 22, 2010; revised on February 11, 2011; accepted on March 7, 2011

1 INTRODUCTION

MicroRNAs (miRNAs) are short (∼22 nt), non-coding RNAs which generally negatively regulate gene expression (Bartel, 2004; He and Hannon, 2004). These short RNAs were first studied in *Caenorhabditis elegans* (Lee *et al.*, 1993; Wightman *et al.*, 1991) and *Drosophila melanogaster* (Lai and Posakony, 1997), but the term miRNA was coined later (Lagos-Quintana *et al.*, 2001). Since their discovery, miRNAs have been identified in a number of animals, plants and viruses (Lagos-Quintana *et al.*, 2002; Molnar *et al.*, 2007; Pasquinelli *et al.*, 2000; Pfeffer *et al.*, 2004; Reinhart *et al.*, 2002; Wang,J.F. *et al.*, 2004). In plants and animals, miRNAs regulate many biological processes (Carrington and Ambros, 2003; Stefani and Slack, 2008), while virus miRNAs mainly target host mRNAs involved in antiviral response, cell proliferation and apoptosis (Pfeffer *et al.*, 2004).

Plant miRNAs are key regulators of a variety of developmental processes, including root development (Boualem *et al.*, 2008; Guo *et al.*, 2005; Wang *et al.*, 2005), vascular development (Kim *et al.*, 2005; Zhou,G.K. *et al.*, 2007), flower development (Chen, 2004; Nag *et al.*, 2009; Wang *et al.*, 2009), leaf morphogenesis (Palatnik *et al.*, 2003) and phase change from vegetative to reproductive development (Lauter *et al.*, 2005). In addition, plant miRNAs are crucial for controlling stress responses as well as nutrient homeostasis (Fujii *et al.*, 2005; Hsieh *et al.*, 2009; Navarro *et al.*, 2006; Sunkar *et al.*, 2006).

In plants, a miRNA gene is commonly transcribed by RNA polymerase II (Lee *et al.*, 2004), giving rise to a primary miRNA transcript (pri-miRNA), which folds into a stem–loop structure (Pasquinelli *et al.*, 2000). Next, mainly unpaired nucleotides at the $3'$ and $5'$ ends are cleaved by a multiprotein complex, resulting in a pre-miRNA (Han *et al.*, 2004; Hiraguri *et al.*, 2005; Kim *et al.*, 2008; Lobbes *et al.*, 2006). Pre-miRNAs feature a huge diversity of lengths (Axtell and Bowman, 2008). Further processing of the pre-miRNA generates a miRNA:miRNA* duplex structure composed of the mature miRNA and its complementary part, the miRNA* (Grishok *et al.*, 2001; Kurihara and Watanabe, 2004). This duplex exhibits high complementarity, including the standard Watson–Crick as well as G:U wobble base pairs, with generally less than four mismatches, and rare and small asymmetric bulges (Meyers *et al.*, 2008). The miRNA:miRNA* duplex is either exported to the cytoplasm, where it is unwound, or it is unwound in the nucleus and the mature miRNA is exported to the cytoplasm (Park *et al.*, 2005). After incorporation into the RNA-induced silencing complex (RISC), the mature miRNA guides the recognition of the target mRNA, which then leads to post-transcriptional regulation of this mRNA (Hutvagner and Zamore, 2002; Vaucheret *et al.*, 2004). Featuring standard and wobble base pairs, with usually less than four mismatches and few bulges, the miRNA is highly complementary to its target mRNA. This way, plant miRNAs generally guide mRNA cleavage (Rhoades *et al.*, 2002). The minimum length of mature miRNAs in plants is 16 nt (Llave *et al.*, 2002).

Some primary transcripts of miRNAs contain introns (Sunkar *et al.*, 2005; Xie *et al.*, 2005; Zhang *et al.*, 2009). In most published cases, the characteristic stem–loop structure is encoded by one of the exons of the miRNA gene. Recently, however, intron-containing miRNA genes have been identified for which the stem–loop structure cannot be predicted based on the unspliced primary transcript (hereafter referred to as pri^{U}-miRNA) (Kutter *et al.*, 2007; Lu *et al.*, 2008; Sunkar *et al.*, 2005). Several of these spliced miRNAs belong to the miR444 family which has only been identified in

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[‡]Present address: Max Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany.

Poaceae (grasses) (Lu *et al.*, 2008; Sunkar *et al.*, 2005). Six members of the miR444 family have been detected in *Oryza sativa* L.ssp. *japonica* (*cv. Nipponbare*) (rice), termed miR444a-f. The primary transcripts of miR444a-f contain an intron and only after this has been removed, the typical hairpin structure can be recognized. In all cases, the exon–exon junction is located in or close to the loop of the stem–loop structure of the pre-miRNA (Lu *et al.*, 2008). The members of the miR444 family in rice target the mRNAs of four MADS-box genes, namely *OsMADS23*, *OsMADS27*, *OsMADS57* and *OsMADS61* (Li *et al.*, 2010; Lu *et al.*, 2008; Sunkar *et al.*, 2005). Most introns in miRNAs feature the canonical splice motif GU-AG (Sunkar *et al.*, 2005; Xie *et al.*, 2005; Zhang *et al.*, 2009).

The prediction of miRNA genes in plants is considerably more difficult than in animals due to a greater heterogeneity of stem–loop structures of plant pre-miRNAs (Axtell and Bowman, 2008; Mendes *et al.*, 2009). Hence, fewer algorithms exist for the prediction of miRNAs in plants than in animals. The available approaches can be subdivided into homology-based searches and *ab initio* methods, where the latter ones also allow the prediction of non-conserved miRNAs (Mendes *et al.*, 2009). Current methods for the *ab initio* prediction of miRNA genes in plants depend on expressed sequence tag (EST) sequences (Jin *et al.*, 2008; Wen *et al.*, 2008) or rely on the predictability of the stem–loop structure from genomic sequence (Adai *et al.*, 2005; Bonnet *et al.*, 2004; Jones-Rhoades and Bartel, 2004; Kadri *et al.*, 2009; Lindow *et al.*, 2007; Teune and Steger, 2010; Wang,X.J. *et al.*, 2004; Xue *et al.*, 2005).

Several tools exist for the prediction and assessment of the interaction between miRNAs and their target mRNAs in plants (Bonnet *et al.*, 2010; Rehmsmeier *et al.*, 2004; Xie and Zhang, 2010; Zhang, 2005). All programs require the mature miRNA as input and consider a high degree of complementarity between the miRNA and its target mRNA with few mismatches and bulges as criterion for interaction of a miRNA and an mRNA. Some programs additionally use criteria such as the estimated minimum free energy of the interaction between the miRNA and its target mRNA (Bonnet *et al.*, 2010; Rehmsmeier *et al.*, 2004) or the evolutionary conservation of this interaction (Zhang, 2005).

Transcripts of miRNAs are underrepresented in EST databases due to their short length and potentially low expression rate. This fact, and the difficulties with the computational detection of spliced miRNAs for which the typical hairpin structure cannot be predicted from genomic sequence, provoke the question of how many of these key regulators remain unidentified in plants.

Here, we present the first method which is able to predict plant pre-miRNAs encoded by intron-containing genes for a given potential target mRNA. Our method SplamiR (Spliced plant miRNAs), combines the well-established bioinformatic tools BLAST (Altschul *et al.*, 1990), GeneSplicer (Pertea *et al.*, 2001) and miR-abela (Sewer *et al.*, 2005) to identify candidate pre-miRNAs in whole genomes whose mature miRNAs likely regulate an mRNA of interest. SplamiR consists of two phases; in Phase 1, sequence pairs with a high degree of complementarity are identified from whole genome or chromosome sequences; and in Phase 2, sequences with complementarity to a given potential target mRNA are searched among those sequence pairs, possible splicing events are considered, and secondary-structure characteristics are evaluated to identify candidate pre-miRNAs. We apply SplamiR to the genomes of rice and *Zea mays* ssp. *mays* L. (maize).

2 METHODS

2.1 Embedded bioinformatic tools

RNA structures cannot only exhibit the standard Watson–Crick base pairs, but also G:U wobble base pairs (Varani and McClain, 2000). To enable BLAST (Altschul *et al.*, 1990) to identify sequence pairs on genomic level which might encode for RNAs forming those stem–loop structures, we modified BLAST internals. Sequence pairs with high complementarity are detected by BLAST by generating the reverse complement and finding a sequence with high similarity to this reverse complement (Fig. 1). This approach implies, that sequence pairs encoding for a G:U wobble base pair are found by BLAST when comparing G to A (as A is the complement of U/T) or T to C (as C is the complement of G). Thus, the scores for the comparisons of G to A and T to C were increased from −3 to 0 in the internal nucleotide scoring matrix of BLAST. We refer to this modified version of BLAST as GU-BLAST.

The interaction of the mature miRNA with its target mRNA can also include G:U wobble base pairs (Allen *et al.*, 2005). Thus, GU-BLAST was also used to find complementary sequences to a given potential target mRNA.

GeneSplicer (Pertea *et al.*, 2001) predicts splice sites with a combination of Markov modeling and decision tree techniques based on training data for true and false splice sites. For this reason, to identify potential splice sites, GeneSplicer was incorporated into SplamiR. The predictions were performed using the provided rice training data which do not include the genes encoding for spliced miRNAs and an acceptance threshold of −30 for acceptor and donor sites. The web server of the miR-abela classifier system (Sewer *et al.*, 2005) was used with default settings to identify candidate pre-miRNAs among generated putative pre-miRNAs. This program distinguishes robust stems of miRNAs from other RNAs with a support vector machine using features such as length, number and nature of bulges, and composition of bases and base pairs of stem–loop structures.

2.2 Algorithm

2.2.1 Phase 1—generation of a database of complementary sequence pairs Pre-miRNAs are characterized by a stem–loop structure (Meyers *et al.*, 2008). In a genome such stem–loop structures are encoded by a stretch of DNA which comprises a sequence pair of two highly complementary sequences.

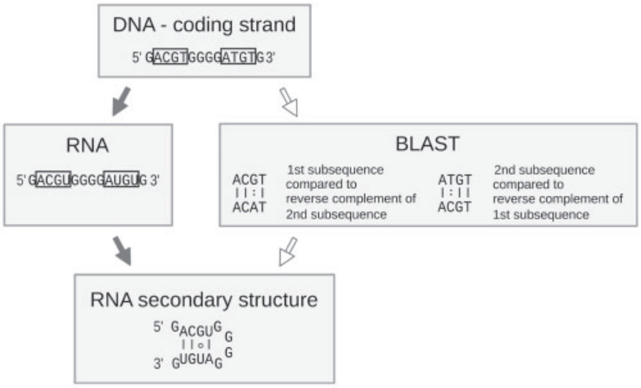

Fig. 1. DNA sequence with complementary sequence pair (boxed) which might encode for an RNA folding into a stem–loop structure containing a G:U base pair and handling of this sequence pair using BLAST (Altschul *et al.*, 1990). The two possible search results with which BLAST would identify the sequence pair encoding for the stem are illustrated on the right. The comparison of G to A and of T to C corresponds to the G:U base pair when the first and the second sequence of the pair, respectively, is used as search sequence.

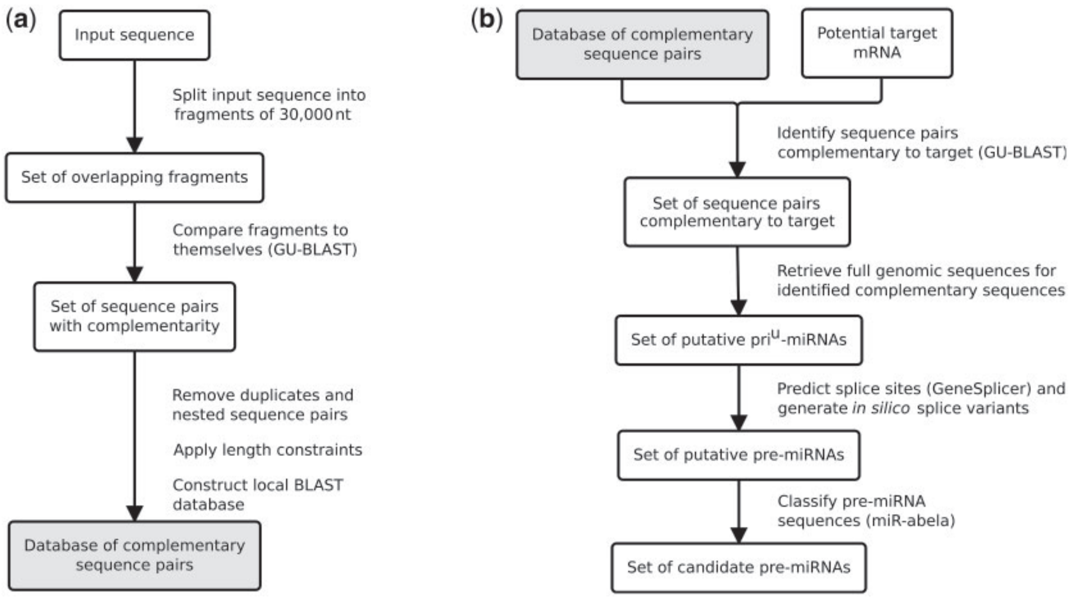
(a)
Input sequence
Split input sequence into fragments of 30,000 nt
Set of overlapping fragments
Compare fragments to themselves (GU-BLAST)
Set of sequence pairs with complementarity
Remove duplicates and nested sequence pairs
Apply length constraints
Construct local BLAST database
Database of complementary sequence pairs
(b)
Database of complementary sequence pairs
Potential target mRNA
Identify sequence pairs complementary to target (GU-BLAST)
Set of sequence pairs complementary to target
Retrieve full genomic sequences for identified complementary sequences
Set of putative pri^U-miRNAs
Predict splice sites (GeneSplicer) and generate *in silico* splice variants
Set of putative pre-miRNAs
Classify pre-miRNA sequences (miR-abela)
Set of candidate pre-miRNAs

Fig. 2. The two phases of SplamiR. In Phase 1 (**a**), a database of complementary sequence pairs (shaded) is created which is then used as basis for the search of candidate pre-miRNAs in Phase 2 (**b**).

Our method takes a sequence which can be as large as whole chromosomes or genomes, as input and first splits it into fragments of 30 000 nt that overlap by 20 000 nt (Fig. 2a). The resulting fragments are then compared with each other with our GU-BLAST using a word length of 9. For each comparison, results of the GU-BLAST search are stored if they are in *plus/minus*-direction. Overlapping hits and BLAST results with a length of <16 nt and >500 nt are omitted. Each sequence of the remaining sequence pairs is stored separately and contains a pointer to the complementary sequence. A local BLAST database of these sequence pairs is created (Fig. 2a). This database can be used in Phase 2 to be searched with a given potential target mRNA, as well as to identify new members of a known miRNA family and for comprehensive genome-wide screenings of all sequence pairs.

2.2.2 Phase 2—identification of candidate pre-miRNAs The entire mature miRNA exhibits high complementarity to its target mRNAs in plants (Rhoades *et al.*, 2002). This information is used in Phase 2 of our method SplamiR. The sequence of a potential target mRNA is used as input (Fig. 2b) and is compared with the database of complementary sequence pairs using our GU-BLAST with a word length of 9. This way, the number of considered complementary sequence pairs is greatly reduced as compared with the whole database. The *e*-values of this GU-BLAST search are stored to later rank candidate pre-miRNAs. For each of the resulting sequence pairs, the genomic sequences are retrieved, ranging from 65 nt upstream of the 5′ end of the 5′-most sequence of the pair, to 65 nt downstream of the 3′ end of the 3′-most sequence of the pair (Fig. 2b and Supplementary Fig. S1). Representing a set of putative pri^U-miRNAs, these sequences are subsequently scanned for canonical splice sites using GeneSplicer. Each pair of thereby predicted splice acceptors and splice donors is considered to frame an intron. Thus, multiple *in silico* splice variants are generated for each genomic sequence (Fig. 2b and Supplementary Fig. S1). All of these variants as well as the original genomic sequences form a set of putative pre-miRNAs, which are then classified by miR-abela to verify whether they represent candidate pre-miRNAs (Fig. 2b). Due to the fact that miR-abela only accepts input sequences of a length of up to 1000 nt, longer putative pre-miRNAs are not considered.

2.3 Datasets

For evaluation, we first created a dataset of artificial spliced miRNAs. To do so, we used 72 known, experimentally verified and not spliced pre-miRNAs from rice which are correctly detectable as pre-miRNAs by miR-abela. Into all of these pre-miRNAs, we introduced a known intron from rice as provided from the The Institute for Genomic Research (TIGR) (Ouyang *et al.*, 2007) at a random position around the middle of each of the sequences. This way, we created 20 different constructs for every pre-miRNA, all with a different intron. Thus, a dataset of 1440 artificial pri^U-miRNAs was generated. Each of these artificial pri^U-miRNAs was put into a fragment of random DNA sequence such that the resulting constructs had a length of 30 000 nt.

We collected experimentally verified and computationally predicted target mRNAs from the literature (Huang *et al.*, 2009; Jian *et al.*, 2010; Jones-Rhoades *et al.*, 2006; Lacombe *et al.*, 2008; Li *et al.*, 2010; Liu *et al.*, 2005; Lu *et al.*, 2008; Luo *et al.*, 2006; Sunkar *et al.*, 2005, 2008; Wu *et al.*, 2009; Xue *et al.*, 2009; Zhou *et al.*, 2010; Zhu *et al.*, 2008) for our dataset of 72 miRNAs. Using SplamiR, every construct containing an artificial pri^U-miRNA was searched with each of the target mRNAs for the inserted miRNA resulting in 4280 tests.

To evaluate SplamiR on known spliced miRNAs, we collected the sequences of the pre-miRNAs of the miR444 family from rice (accession numbers MI0001719, MI0006974-MI0006978) from miRBase (Griffiths-Jones, 2004). The exon–intron boundaries and the sequence of the introns were determined by alignment of the pre-miRNAs to the genomic sequence of *O.sativa* ssp. *japonica*, release 6.1, which was retrieved from TIGR (Ouyang *et al.*, 2007). The mRNA sequences of the predicted target genes of the members of the miR444 family, *OsMADS23*, *OsMADS27*, *OsMADS57* and *OsMADS61* were also obtained from TIGR, where only the primary splice form was considered. In case of *OsMADS61*, we assume that the mRNA sequence is more likely a concatenation of the sequences given with the identifiers LOC_Os04g38770 and LOC_Os04g38780.

To identify new pre-miRNAs, we retrieved the genome sequence of *Z. mays* ssp. *mays* L. (Schnable *et al.*, 2009) from http://maizesequence.org. A potential target gene of a member of the miR444 family in maize is *ZmMADS2* (Schreiber *et al.*, 2004). The mRNA sequence of *ZmMADS2* (accession number AF112149) and an EST sequence (accession number: EE182122) corresponding to the pre-miRNA were obtained from the National Centre for Biotechnology Information (NCBI) (Benson *et al.*, 2010).

3 RESULTS AND DISCUSSION

3.1 Rationale for integrated tools and applied constraints

The presented method for the prediction of spliced miRNAs makes use of two crucial characteristics of plant miRNAs. First, the high complementarity between the mature miRNA and the miRNA* in the duplex structure of the pre-miRNA (Meyers *et al.*, 2008) is the reasoning behind the construction of a database of highly complementary sequence pairs in Phase 1 of SplamiR. Secondly, miRNAs are highly complementary to their target mRNA (Rhoades *et al.*, 2002), which is used as criterion for Phase 2 of our method.

3.1.1 Phase 1 In Phase 1, SplamiR identifies highly complementary sequence pairs within the input sequence using our GU-BLAST. By altering the internal nucleotide scoring matrix, the *e*-value is lowered for the comparison of complementary sequence pairs which might encode for stem–loop structures including G:U wobble base pairs. Thus, more sequences are identified and resulting sequences are likely longer than those for which G:U wobble base pairs are not considered.

BLAST (Altschul *et al.*, 1990) was chosen for this task as it enables fast comparison of sequences. To ensure fast prediction, secondary structures are not computed at this stage but are used for classification when applying miR-abela (Sewer *et al.*, 2005) at the end of Phase 2 of our method.

Identification of an excessive number of complementary sequence pairs representing low complexity regions in this step is prevented by applying the build-in filters of BLAST. In contrast, transposable elements were purposely not excluded as some miRNAs evolved from these elements (Piriyapongsa and Jordan, 2008; Piriyapongsa *et al.*, 2007).

To further speed up the search, the input sequence is split into fragments of 30 000 nt which overlap by 20 000 nt (Fig. 2a). By choosing these fragment and overlap sizes, it is ensured that sequence pairs can be identified which are as much as 20 000 nt apart. When splicing is considered in Phase 2, the sequence in between the complementary sequence pairs might be predicted to represent an intron. Thus, pri^{U}-miRNAs which contain an intron with a length of nearly 20 000 nt are predictable by our method. Intronic sequences in plants usually range from 60 to 10 000 nt (Lorkovic *et al.*, 2000). The members of the miR444 family have introns with lengths between 1189 to 6721 nt (Lu *et al.*, 2008; Sunkar *et al.*, 2005). Accordingly, our method is likely able to detect most spliced miRNAs in plants.

If a genomic region identified by the GU-BLAST search encodes for a pre-miRNA, then the length of the result approximately corresponds to the length of one arm of its stem–loop structure, and thus represents roughly half of the length of its complete structure. The results of the GU-BLAST search are stored if they have lengths between 16 and 500 nt. The minimum length of 16 nt reflects the minimum length of the miRNA:miRNA* duplex (Llave *et al.*, 2002). The maximum length of 500 nt was chosen as the maximal length of putative pre-miRNA sequences which can be tested by miR-abela is 1000 nt, and lengths of currently stored pre-miRNAs of flowering plants (Magnoliophyta) in miRBase (Griffiths-Jones, 2004) range between 54 and 734 nt. Thus, our minimal and maximal values of 16 and 500 nt, respectively, for one arm of the stem–loop sequence seem reasonable.

3.1.2 Phase 2 In Phase 2 of our method, candidate pre-miRNAs which might regulate a given potential mRNA are identified (Fig. 2b). To do so, GU-BLAST is used to find sequences in the database of complementary sequence pairs which are also complementary to the given input mRNA. Again, for fast prediction, minimum free energy or other constraints on the degree of complementarity as deduced from experimental findings are not considered at this stage. Available target prediction tools were also not incorporated as they require the sequence of a mature miRNA as input (Bonnet *et al.*, 2010; Rehmsmeier *et al.*, 2004; Xie and Zhang, 2010; Zhang, 2005). The mature miRNA cannot be inferred from predicted pre-miRNAs. However, we suggest an application of these tools on candidate miRNAs predicted by our method. The region with complementarity to the target gene of each candidate pre-miRNA could be split into fragments with a length of ~20 nt and these fragments could then be used as input for available target prediction tools.

As we aim at the prediction of spliced miRNAs, the exon–intron boundaries of the retrieved putative pri^{U}-miRNAs need to be determined. Different gene prediction tools (Brejova *et al.*, 2005; Korf, 2004; Majoros *et al.*, 2004) failed to detect the exon–intron structure for our test cases. However, GeneSplicer (Pertea *et al.*, 2001) appeared to be suitable for the detection of potential splice sites in putative pri^{U}-miRNAs.

GeneSplicer only detects canonical GU and AG splicing motifs. As known introns in miRNA genes generally feature these canonical splicing motifs (Sunkar *et al.*, 2005; Xie *et al.*, 2005; Zhang *et al.*, 2009), the use of GeneSplicer does not seem to impose serious restrictions.

For the detection of splicing motifs, GeneSplicer uses a sliding window of 80 nt. To allow the detection of splicing motifs with a distance of just >15 nt from the 5′ end of the upstream-most sequence or from the 3′ end of the downstream-most sequence of the complementary sequence pair, 65 nt genomic sequence were added to the 5′ end and the 3′ end of the complementary sequence pairs (Supplementary Fig. S1).

For each putative pri^{U}-miRNA, all possible combinations of predicted splice donors and acceptors are used to create *in silico* splice variants. Each splice variant is tested whether it represents a candidate pre-miRNA. The full genomic sequence is also considered to allow the detection of pre-miRNAs for which the corresponding miRNA genes do not contain introns.

Pertaining to our method, there are four different relevant exon–intron structures in primary transcripts of miRNAs (Supplementary Fig. S2). First, after removal of introns, the pre-miRNA folds into a structure in which exon–exon junctions are located between one of the ends of the structure and the miRNA or the miRNA* (section I in Supplementary Fig. S2a). In these most common cases, removal of the intron is not necessary for the prediction of the stem–loop structure (Xie *et al.*, 2005; Zhang *et al.*, 2009). These pre-miRNAs can be detected by both, standard miRNA-prediction programs and SplamiR. Secondly, there are cases where the exon–exon junction is located in the part of the structure which is enclosed by the miRNA and the miRNA* (section III in Supplementary Fig. S2b) (Lu *et al.*, 2008; Sunkar *et al.*, 2005). These cases are covered by SplamiR. Thirdly, it is possible, that pri^{U}-miRNAs contain multiple introns for which the exon–exon junctions are located in the part of the pre-miRNA structure which is enclosed by the miRNA and the miRNA* (section III in Supplementary Fig. S2c). Even though no such pre-miRNAs are known so far, our method is likely able to detect those pre-miRNAs. In this case, the splice donor of the upstream-most intron and the splice acceptor of the downstream-most intron are considered as flanking a single intron and an *in silico* splice variant is created which is similar to the actual pre-miRNA but has a shorter loop section. This putative pre-miRNA likely passes the succeeding classification step as the shorter loop section presumably does not negatively influence its classification. Finally, an exon–exon junction could be located in the part of the structure giving rise to the miRNA:miRNA* duplex (section II in Supplementary Fig. S2d). Our method is not able to detect these pre-miRNAs in case the remaining sequence complementarity is insufficient for identification. However, no such pre-miRNAs are known so far.

Only a few programs exist for the classification of plant miRNAs (Adai *et al.*, 2005; Bonnet *et al.*, 2004; Jones-Rhoades and Bartel, 2004; Lindow and Krogh, 2005; Wang,X.J. *et al.*, 2004). These programs are not suitable to be incorporated into SplamiR as they only predict conserved miRNAs or have a low specificity. Thus, classification programs trained for animal miRNAs (Hertel and Stadler, 2006; Lai *et al.*, 2003; Lim *et al.*, 2003; Sewer *et al.*, 2005) were tested. On a dataset of nearly 4000 randomly chosen RNA sequences (excluding miRNAs) of flowering plants (Magnoliophyta) taken from Rfam (Gardner *et al.*, 2009) and 600 randomly chosen known pre-miRNAs of flowering plants listed in miRBase, miR-abela yielded a sensitivity and a specificity of 99.8 and 58.8%, respectively. Hence, miR-abela was incorporated into SplamiR. Even though SplamiR's detection rate is good using miR-abela, it can be easily increased once algorithms become available, which allow better classification of plant pre-miRNAs, by substituting miR-abela by these new algorithms.

Like most other methods for the prediction of miRNAs (Bonnet *et al.*, 2004; Hertel and Stadler, 2006; Lim *et al.*, 2003; Lindow and Krogh, 2005), our method predicts pre-miRNAs. The exact position of the mature miRNA cannot be determined in all cases, especially in cases where more than one mature miRNA is excisable from one pre-miRNA (Lu *et al.*, 2008; Molnar *et al.*, 2007; Xie *et al.*, 2005). However, in many cases, the position of the mature miRNA can be confined as the mature miRNA is highly complementary to the target mRNA (Rhoades *et al.*, 2002).

3.1.3 Time and memory usage As indicated above, the integrated tools ensure fast performance of SplamiR. On a 2.5 GHz single core system processing of 1 MB of genomic sequence in Phase 1 takes 13 s on average. The approximate computation time for searching the database with a target mRNA, generation of splice variants and classification of putative pre-miRNAs is 2 h per genome. Thereby, the computation time is heavily dependent on the number of splice variants. Generation and classification of one splice variant takes ~14 s. Memory usage is low, <256 MB, in both phases of SplamiR. Hence, pre-miRNAs can be predicted on whole genome sequences in modest time using little memory with our method.

3.2 Simulation studies

In order to evaluate the performance of SplamiR, we conducted simulation studies on a dataset of artificial pri^{U}-miRNAs and their corresponding target mRNAs. The artificial pri^{U}-miRNAs were generated by combining each of 72 known, not spliced pre-miRNAs from rice with 20 known rice introns. These constructs, as well as target mRNAs of the 72 miRNAs as identified from the literature, were used as input for SplamiR for our simulation studies.

3.2.1 Phase 1 Searching our dataset of 1440 constructs containing artificial pri^{U}-miRNAs resulted in 943 cases with the identification of the correct sequence pair corresponding to the inserted pre-miRNA (±50 nt). This is equivalent to a reasonable sensitivity of 65.5% of Phase 1 of SplamiR. In total, 2558 complementary sequence pairs were detected, of which 60.6% are false positive results. Hence, prediction of spliced miRNAs without prior knowledge of target mRNAs may be challenging with SplamiR as all false complementary sequence pairs would also be considered in Phase 2. In contrast, searching the database using a target mRNA greatly reduces the number of considered false positive sequence pairs as discussed below.

3.2.2 Phase 2 We conducted 4280 tests searching all of the 1440 constructs with each target mRNA of the inserted miRNA. In 2790 of these tests a sequence pair matching approximately to the inserted pre-miRNA (±100 nt) could be identified. This equates to a sensitivity of 65.2% of this step of SplamiR. In total, 5086 sequence pairs with complementarity to the target mRNA were detected. Only 43 (0.8%) of these sequence pairs do not coincide with the inserted pre-miRNA and represent false positive results.

A total of 17 392 splice variants was created, i.e. an average of 3.4 splice variants per identified pri^{U}-miRNA. Overall, 1454 of the created splice variants were classified positively by miR-abela (Sewer *et al.*, 2005). Once again, the number of false positives was low with only 26 (1.8%) positively classified pre-miRNAs which do not correspond to an inserted miRNA. In 910 cases of the 4280 interactions tested a splice variant coinciding with the inserted pre-miRNA (±100 nt) was positively classified as candidate pre-miRNA by miR-abela. This equates to a sensitivity of 21.3% for Phase 2 of SplamiR.

There is a strong decrease from 2790 tests for which the correct complementary sequence pair was identified from the database to 910 tests for which a valid pre-miRNA was detected. This reduction is caused by application of GeneSplicer (Pertea *et al.*, 2001) and miR-abela (Sewer *et al.*, 2005). Our artificial pri^{U}-miRNAs are composed of known introns from rice and known pre-miRNAs which are split into two artificial exons. It is possible that GeneSplicer does not recognize correct splice sites in our dataset as our artificial exons likely miss important recognition sites. Hence, sensitivity of SplamiR might be even higher on genuine spliced pre-miRNAs than the result of the simulation studies suggest.

3.3 Test case: miR444 family in rice

The members of the miR444 family contain introns in a way that the characteristic stem–loop structure could previously not be predicted from genomic sequences (Lu *et al.*, 2008; Sunkar *et al.*, 2005). Consequently, to evaluate our method on known spliced miRNAs, it was applied to the 12 chromosomes of rice in combination with

Table 1. SplamiR results for the target mRNAs of the miR444 family

Target mRNA	Putative pri^U-miRNAs	Putative pre-miRNAs	Candidate pre-miRNAs	Unique loci
OsMADS23	54	318	37	16
OsMADS27	110	601	57	17
OsMADS57	87	577	126	25
OsMADS61	177	785	131	34

Number of sequences handled at each step in Phase 2 of SplamiR when using the four target mRNAs of the miR444 family of miRNAs as input sequences. Putative pre-miRNAs are *in silico* splice variants which have not been tested using miR-abela yet. *In silico* splice variants which have been classified as pre-miRNAs by miR-abela are termed candidate pre-miRNAs. Details about unique loci are shown in Supplementary Table S1.

the target mRNAs of the members of the miR444 family, namely *OsMADS23*, *OsMADS27*, *OsMADS57* and *OsMADS61* (Li *et al.*, 2010; Lu *et al.*, 2008; Sunkar *et al.*, 2005).

In Phase 1, 730 146 complementary sequence pairs were found in rice and stored in our database. Among these sequence pairs are sequences of all six miR444-family members. When querying this database with the target mRNA *OsMADS23*, 54 sequence pairs are found for which one of the two subsequences is highly complementary to this input mRNA (Table 1). Again, sequences of all six miR444-family members are among the putative pri^U-miRNAs.

Scanning the corresponding genomic regions of all 54 putative pri^U-miRNAs using GeneSplicer and generating all possible *in silico* splice variants, results in a total of 318 putative pre-miRNAs with a length of <1000 nt (Table 1). In case of the six miR444-family members, GeneSplicer is able to predict the correct donor and acceptor splice sites. However, the correct splice site is not always the one with the best score in GeneSplicer. For example, the score for the real donor site of miR444a was 6.73 and thus decisively worse than the score of 0.03 for a predicted donor site in the same sequence. The number of splice sites ranges from one donor and two acceptor sites (miR444e and miR444f) to five donor and five acceptor sites (miR444a). Applying this step to all pri^U-miRNAs of the six miR444-family members, results in 15 *in silico* splice variants <1000 nt for miR444a, 9 variants for miR444b, 13 variants for miR444c, 8 variants for miR444d, 2 variants for miR444e and 2 variants for miR444f.

Classifying all the 318 generated *in silico* splice variants using miR-abela, results in 37 candidate pre-miRNAs (Table 1 and Supplementary Table S1). When neglecting variants, which differ slightly in their splice sites and/or in their starting or ending positions in the genome, a set of 16 candidate pre-miRNAs remains. This set includes all six miR444-family members and 10 additional, unknown loci which are neither present in miRBase nor show similarity to any other mature miRNA sequence. Hence, miR-abela is able to correctly classify the miR444-family members as pre-miRNAs, even if the 5′ and/or 3′ ends of the putative pre-miRNA sequences differ slightly from the verified sequences.

When using the mRNAs of *OsMADS27*, *OsMADS57* and *OsMADS61* as input sequences for Phase 2 of our method, sets of 17, 25 and 34 candidate pre-miRNAs, respectively, are generated (Table 1 and Supplementary Table S1). As in the case of *OsMADS23*, also for *OsMADS57* all six miR444-family members are identified by our method, which is in accordance with previous predictions (Lu *et al.*, 2008). For *OsMADS27*, miR444e is the only miR444-family member which is not detected by our method. *OsMADS27* was predicted to be regulated by all six miR444-family members (Lu *et al.*, 2008). However, this regulation has not been verified yet and might not occur *in vivo*. *OsMADS61* was predicted to be only regulated by miR444b and miR444c (Lu *et al.*, 2008). With our method, additionally miR444a and miR444d are identified as candidate pre-miRNAs when using the mRNA of *OsMADS61* as input sequence. Both, miR444a and miR444d contain sequences that are highly complementary to the mRNA of *OsMADS61*, with only three mismatches, and thus might indeed regulate *OsMADS61*. In accordance with the predictions (Lu *et al.*, 2008), miR444e and miR444f are not identified by our method when using the mRNA of *OsMADS61* as input.

For each candidate pre-miRNA, we retrieved the *e*-value of the GU-BLAST search identifying the corresponding complementary sequence pair using the target mRNA as input. This *e*-value is very low, <0.01, for all miR444-family members (Supplementary Table S1). Hence, we introduced a ranking based on this *e*-value representing the complementarity of the candidate pre-miRNA to the target mRNA. We suggest that candidate pre-miRNAs having an accordant *e*-value of less than 1 could be considered as high priority candidates.

Due to the complementarity of the two strands of DNA and the complementarity within pre-miRNAs (resulting in the typical stem–loop structure which is crucial for their recognition), it is not always possible to distinguish from which strand of DNA a miRNA gene is transcribed (Adai *et al.*, 2005; Lai *et al.*, 2003; Sandmann and Cohen, 2007). Nevertheless, our method is able to identify the coding strand of the genes of miR444a, miR444b, miR444e and miR444f. For miR444c and miR444d, our method predicts a candidate pre-miRNA for both strands. In these cases, promoter detection methods (Abeel *et al.*, 2009; Zhou,X. *et al.*, 2007) might help to determine from which strand the miRNA gene is actually transcribed.

In total, SplamiR detects 79 candidate pre-miRNAs for the four mRNAs of *OsMADS23*, *OsMADS27*, *OsMADS57* and *OsMADS61*. The largest number of candidate pre-miRNAs is identified when using the mRNA of *OsMADS61* as input (34 candidate pre-miRNAs, including four miR444-family members, Table 1 and Supplementary Table S1). Of the 79 candidate pre-miRNAs, six represent members of the miR444 family of miRNAs while some of the remaining 73 candidate pre-miRNAs might be false positive results. To reduce the number of remaining candidate pre-miRNAs, their pre-miRNA structures can be studied in more detail to reveal violations of properties that are not checked by miR-abela, like mismatches between the mature miRNA and the target mRNA surrounding the assumed site of mRNA cleavage (Schwab *et al.*, 2005).

3.4 Predictions of miRNAs in maize

For maize, the expression of one member of the miR444 family has been verified by RNA gel blots (Lu *et al.*, 2008; Sunkar *et al.*, 2005). A recent study on the characterization of miRNA genes in the maize genome failed to identify the genomic locus encoding this miR444 due to the presence of a large intron in the pri^U-miRNA (Zhang *et al.*, 2009).

To find genes encoding miR444-family members, we employed our method to the 10 chromosomes of maize. SplamiR detects a total

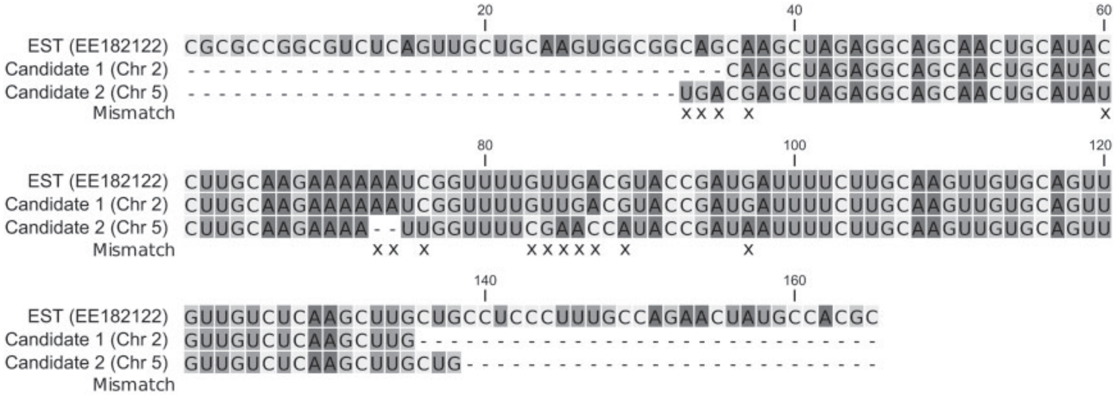
20
40
60
EST (EE182122) CGCGCCGGCGUCUCAGUUGCUGCAAGUGGCGGCAGCAAGCUAGAGGCAGCAACUGCAUAC
Candidate 1 (Chr 2) -----------------------------------CAAGCUAGAGGCAGCAACUGCAUAC
Candidate 2 (Chr 5) --------------------------------UGACGAGCUAGAGGCAGCAACUGCAUAU
Mismatch XXX X X
80
100
120
EST (EE182122) CUUGCAAGAAAAAAUCGGUUUUGUUGACGUACCGAUGAUUUUCUUGCAAGUUGUGCAGUU
Candidate 1 (Chr 2) CUUGCAAGAAAAAAUCGGUUUUGUUGACGUACCGAUGAUUUUCUUGCAAGUUGUGCAGUU
Candidate 2 (Chr 5) CUUGCAAGAAAA--UUGGUUUUCGAACCAUACCGAUAAUUUUCUUGCAAGUUGUGCAGUU
Mismatch XX X XXXXX X X
140
160
EST (EE182122) GUUGUCUCAAGCUUGCUGCCUCCCUUUGCCAGAACUAUGCCACGC
Candidate 1 (Chr 2) GUUGUCUCAAGCUUG------------------------------
Candidate 2 (Chr 5) GUUGUCUCAAGCUUGCUG---------------------------
Mismatch

Fig. 3. Alignment of an EST of the miR444-family member in maize and the two candidate pre-miRNAs identified by SplamiR. The candidate pre-miRNA found on chromosome 2 has an identical sequence compared with the EST, while the candidate pre-miRNA found on chromosome 5 has 13 mismatches and two deletions (all differences are indicated by X) compared with the EST sequence in the alignable region.

of 4 495 838 complementary sequence pairs in the maize genome. Using the mRNA of the maize gene *ZmMADS2*, which is closely related to the rice target genes considered above (Arora *et al.*, 2007; Becker and Theissen, 2003; Schreiber *et al.*, 2004), we identified 14 candidate pre-miRNAs (Supplementary Table S1). Considering the comparatively large genome size of maize (Schnable *et al.*, 2009), this represents a reasonably low number of candidate pre-miRNAs. None of the candidate pre-miRNAs are listed in miRBase. Two of the 14 candidate pre-miRNAs exhibit sequence similarity to the pre-miR444 sequences from rice. The remaining 12 candidate pre-miRNAs do not show similarity to any known pre-miRNA. A more detailed comparison revealed identity between partial sequences of two of the candidate pre-miRNAs and the mature miRNA sequences of miR444b and miR444c of rice. These two candidate pre-miRNAs are encoded by genes which are located on the maize chromosomes 2 and 5, respectively.

An EST sequence of the previously verified miR444 from maize (Lu *et al.*, 2008; Sunkar *et al.*, 2005) was retrieved from GenBank (Benson *et al.*, 2010). Both candidate pre-miRNAs are shorter than, but otherwise nearly identical to this EST sequence (Fig. 3). The candidate pre-miRNA on chromosome 2 is identical to the EST sequence in the alignable region. Thus, we identified the genomic location of the gene encoding the known miR444-family member in maize. Analysis of the genomic locus revealed that the gene contains an intron with a length of 3397 nt. A gene with EST support (GRMZM2G032942) is annotated at this genomic locus. The position of the intron of the pri^{U}-miRNA as predicted by our method is in accordance with an intron of the annotated gene. Comparison of the candidate pre-miRNA on chromosome 5 with the EST sequence reveals only 13 mismatches and two deletions in the alignable region (Fig. 3). Consequently, this candidate pre-miRNA likely represents a previously undiscovered member of the miR444 family in maize. Also in this case, a gene with EST support (GRMZM2G016651) is annotated in the corresponding genomic region. The pri^{U}-miRNA contains an intron with a length of 7661 nt, consistent with the annotated gene. The high complementarity of parts of the candidate pre-miRNAs to the potential target mRNA *ZmMADS2* strongly suggests regulation of *ZmMADS2* by these candidate pre-miRNAs.

The remaining 12 candidate pre-miRNAs do not exhibit convincing similarity to known pre-miRNAs and thus await verification.

4 CONCLUSIONS

We presented SplamiR, the first method for the prediction of spliced miRNAs in plants. Our method predicts sequences and structures of pre-miRNAs. SplamiR has a modular structure which provides great flexibility as individual modules can be replaced by more specific ones as soon as these become available.

We demonstrated that SplamiR is able to detect all members of the miR444 family of miRNAs in rice, a family of miRNAs for which the characteristic stem–loop structure cannot be predicted on genomic sequence and which has hence only been identified experimentally. Furthermore, using SplamiR, we have detected a previously unidentified member of the miR444 family in maize.

Our method will help to determine whether an mRNA of interest is regulated by a miRNA and will reveal more of these key regulators which have, so far, escaped identification.

ACKNOWLEDGEMENTS

We are grateful to all members of the Theißen laboratory for helpful discussions and to the Friedrich Schiller University in Jena for general support. We wish to thank three anonymous reviewers for their valuable comments.

Conflict of Interest: none declared.

REFERENCES

Abeel,T. *et al.* (2009) Toward a gold standard for promoter prediction evaluation. *Bioinformatics*, **25**, i313–i320.

Adai,A. *et al.* (2005) Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res.*, **15**, 78–91.

Allen,E. *et al.* (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, **121**, 207–221.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arora,R. *et al.* (2007) MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics*, **8**, 242.

Axtell,M.J. and Bowman,J.L. (2008) Evolution of plant microRNAs and their targets. *Trends Plant Sci.*, **13**, 343–349.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Becker,A. and Theissen,G. (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol Phylogenet. Evol.*, **29**, 464–489.

Benson,D.A. *et al.* (2010) GenBank. *Nucleic Acids Res.*, **38**, D46–D51.

Bonnet,E. *et al.* (2010) TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics*, **26**, 1566–1568.

Bonnet,E. *et al.* (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc. Natl Acad. Sci. USA*, **101**, 11511–11516.

Boualem,A. *et al.* (2008) MicroRNA166 controls root and nodule development in Medicago truncatula. *Plant J.*, **54**, 876–887.

Brejova,B. *et al.* (2005) ExonHunter: a comprehensive approach to gene finding, *Bioinformatics*, **21** (Suppl. 1), i57–i65.

Carrington,J.C. and Ambros,V. (2003) Role of microRNAs in plant and animal development. *Science*, **301**, 336–338.

Chen,X. (2004) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science*, **303**, 2022–2025.

Fujii,H. *et al.* (2005) A miRNA involved in phosphate-starvation response in Arabidopsis. *Curr. Biol.*, **15**, 2038–2043.

Gardner,P.P. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.

Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.

Grishok,A. *et al.* (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, **106**, 23–34.

Guo,H.S. *et al.* (2005) MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for arabidopsis lateral root development. *Plant Cell*, **17**, 1376–1386.

Han,M.H. *et al.* (2004) The Arabidopsis double-stranded RNA-binding protein HYL1 plays a role in microRNA-mediated gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 1093–1098.

He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.

Hertel,J. and Stadler,P.F. (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, **22**, e197–e202.

Hiraguri,A. *et al.* (2005) Specific interactions between Dicer-like proteins and HYL1/DRB-family dsRNA-binding proteins in Arabidopsis thaliana. *Plant Mol. Biol.*, **57**, 173–188.

Hsieh,L.C. *et al.* (2009) Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. *Plant Physiol.*, **151**, 2120–2132.

Huang,S.Q. *et al.* (2009) Heavy metal-regulated new microRNAs from rice. *J. Inorg. Biochem.*, **103**, 282–287.

Hutvagner,G. and Zamore,P.D. (2002) A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, **297**, 2056–2060.

Jian,X. *et al.* (2010) Identification of novel stress-regulated microRNAs from Oryza sativa L. *Genomics*, **95**, 47–55.

Jin,W. *et al.* (2008) Identification and verification of microRNA in wheat (Triticum aestivum). *J. Plant Res.*, **121**, 351–355.

Jones-Rhoades,M.W. and Bartel,D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, **14**, 787–799.

Jones-Rhoades,M.W. *et al.* (2006) MicroRNAS and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, **57**, 19–53.

Kadri,S. *et al.* (2009) HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics*, **10**, S35.

Kim,J. *et al.* (2005) microRNA-directed cleavage of ATHB15 mRNA regulates vascular development in Arabidopsis inflorescence stems. *Plant J.*, **42**, 84–94.

Kim,S. *et al.* (2008) Two cap-binding proteins CBP20 and CBP80 are involved in processing primary microRNAs. *Plant Cell Physiol.*, **49**, 1634–1644.

Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Kurihara,Y. and Watanabe,Y. (2004) Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc. Natl Acad. Sci. USA*, **101**, 12753–12758.

Kutter,C. *et al.* (2007) MicroRNA-mediated regulation of stomatal development in Arabidopsis. *Plant Cell*, **19**, 2417–2429.

Lacombe,S. *et al.* (2008) Identification of precursor transcripts for 6 novel miRNAs expands the diversity on the genomic organisation and expression of miRNA genes in rice. *BMC Plant Biol.*, **8**, 123.

Lagos-Quintana,M. *et al.* (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

Lagos-Quintana,M. *et al.* (2002) Identification of tissue-specific microRNAs from mouse. *Curr. Biol.*, **12**, 735–739.

Lai,E.C. and Posakony,J.W. (1997) The Bearded box, a novel 3′ UTR sequence motif, mediates negative post-transcriptional regulation of bearded and enhancer of split complex gene expression. *Development*, **124**, 4847–4856.

Lai,E.C. *et al.* (2003) Computational identification of Drosophila microRNA genes. *Genome Biol.*, **4**, R42.

Lauter,N. *et al.* (2005) microRNA172 down-regulates glossy15 to promote vegetative phase change in maize. *Proc. Natl Acad. Sci. USA*, **102**, 9412–9417.

Lee,R.C. *et al.* (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

Lee,Y. *et al.* (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, **23**, 4051–4060.

Li,Y.F. *et al.* (2010) Transcriptome-wide identification of microRNA targets in rice. *Plant J.*, **62**, 742–759.

Lim,L.P. *et al.* (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.

Lindow,M. and Krogh,A. (2005) Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics*, **6**, 119.

Lindow,M. *et al.* (2007) Intragenomic matching reveals a huge potential for miRNA-mediated regulation in plants. *PLoS Comput. Biol.*, **3**, e238.

Liu,B. *et al.* (2005) Loss of function of OsDCL1 affects microRNA accumulation and causes developmental defects in rice. *Plant Physiol.*, **139**, 296–305.

Llave,C. *et al.* (2002) Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, **14**, 1605–1619.

Lobbes,D. *et al.* (2006) SERRATE: a new player on the plant microRNA scene. *EMBO Rep.*, **7**, 1052–1058.

Lorkovic,Z.J. *et al.* (2000) Pre-mRNA splicing in higher plants. *Trends Plant Sci.*, **5**, 160–167.

Lu,C. *et al.* (2008) Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). *Proc. Natl Acad. Sci. USA*, **105**, 4951–4956.

Luo,Y.C. *et al.* (2006) Rice embryogenic calli express a unique set of microRNAs, suggesting regulatory roles of microRNAs in plant post-embryogenic development. *FEBS Lett.*, **580**, 5111–5116.

Majoros,W.H. *et al.* (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

Mendes,N.D. *et al.* (2009) Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res.*, **37**, 2419–2433.

Meyers,B.C. *et al.* (2008) Criteria for annotation of plant MicroRNAs. *Plant Cell*, **20**, 3186–3190.

Molnar,A. *et al.* (2007) miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, **447**, 1126–1129.

Nag,A. *et al.* (2009) miR319a targeting of TCP4 is critical for petal growth and development in Arabidopsis. *Proc. Natl Acad. Sci. USA*, **106**, 22534–22539.

Navarro,L. *et al.* (2006) A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science*, **312**, 436–439.

Ouyang,S. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.

Palatnik,J.F. *et al.* (2003) Control of leaf morphogenesis by microRNAs. *Nature*, **425**, 257–263.

Park,M.Y. *et al.* (2005) Nuclear processing and export of microRNAs in Arabidopsis. *Proc. Natl Acad. Sci. USA*, **102**, 3691–3696.

Pasquinelli,A.E. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.

Pertea,M. *et al.* (2001) GeneSplicer: a new computational method for splice site prediction, *Nucleic Acids Res.*, **29**, 1185–1190.

Pfeffer,S. *et al.* (2004) Identification of virus-encoded microRNAs. *Science*, **304**, 734–736.

Piriyapongsa,J. and Jordan,I.K. (2008) Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA*, **14**, 814–821.

Piriyapongsa,J. *et al.* (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics*, **176**, 1323–1337.

Rehmsmeier,M. *et al.* (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Reinhart,B.J. *et al.* (2002) MicroRNAs in plants. *Genes Dev.*, **16**, 1616–1626.

Rhoades,M.W. *et al.* (2002) Prediction of plant microRNA targets. *Cell*, **110**, 513–520.

Sandmann,T. and Cohen,S.M. (2007) Identification of novel Drosophila melanogaster microRNAs. *PLoS One*, **2**, e1265.

Schnable,P.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Schreiber,D.N. *et al.* (2004) The MADS box transcription factor ZmMADS2 is required for anther and pollen maturation in maize and accumulates in apoptotic bodies during anther dehiscence. *Plant Physiol.*, **134**, 1069–1079.

Schwab,R. *et al.* (2005) Specific effects of microRNAs on the plant transcriptome. *Dev. Cell*, **8**, 517–527.

Sewer,A. *et al.* (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267.

Stefani,G. and Slack,F.J. (2008) Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol.*, **9**, 219–230.

Sunkar,R. *et al.* (2005) Cloning and characterization of microRNAs from rice. *Plant Cell*, **17**, 1397–1411.

Sunkar,R. *et al.* (2006) Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. *Plant Cell*, **18**, 2051–2065.

Sunkar,R. *et al.* (2008) Identification of novel and candidate miRNAs in rice by high throughput sequencing. *BMC Plant Biol.*, **8**, 25.

Teune,J.H. and Steger,G. (2010) NOVOMIR: de novo prediction of microRNA-coding regions in a single plant-genome. *J. Nucleic Acids*, **2010**, 10.

Varani,G. and McClain,W.H. (2000) The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep.*, **1**, 18–23.

Vaucheret,H. *et al.* (2004) The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. *Genes Dev.*, **18**, 1187–1197.

Wang,J.F. *et al.* (2004) Identification of 20 microRNAs from Oryza sativa. *Nucleic Acids Res.*, **32**, 1688–1695.

Wang,J.W. *et al.* (2005) Control of root cap formation by MicroRNA-targeted auxin response factors in Arabidopsis. *Plant Cell*, **17**, 2204–2216.

Wang,J.W. *et al.* (2009) miR156-regulated SPL transcription factors define an endogenous flowering pathway in Arabidopsis thaliana. *Cell*, **138**, 738–749.

Wang,X.J. *et al.* (2004) Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biol.*, **5**, R65.

Wen,J. *et al.* (2008) Computational prediction of candidate miRNAs and their targets from Medicago truncatula non-protein-coding transcripts. *In Silico Biol.*, **8**, 291–306.

Wightman,B. *et al.* (1991) Negative regulatory sequences in the lin-14 3′-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes Dev.*, **5**, 1813–1824.

Wu,L. *et al.* (2009) Rice microRNA effector complexes and targets. *Plant Cell*, **21**, 3421–3435.

Xie,F. and Zhang,B. (2010) Target-align: a tool for plant microRNA target identification. *Bioinformatics*, **26**, 3002–3003.

Xie,Z. *et al.* (2005) Expression of Arabidopsis MIRNA genes. *Plant Physiol.*, **138**, 2145–2154.

Xue,C. *et al.* (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine, *BMC Bioinformatics*, **6**, 310.

Xue,L.J. *et al.* (2009) Characterization and expression profiles of miRNAs in rice seeds. *Nucleic Acids Res.*, **37**, 916–930.

Zhang,L. *et al.* (2009) A genome-wide characterization of microRNA genes in maize. *PLoS Genet.*, **5**, e1000716.

Zhang,Y. (2005) miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res.*, **33**, W701–W704.

Zhou,G.K. *et al.* (2007) Overexpression of miR165 affects apical meristem formation, organ polarity establishment and vascular development in Arabidopsis. *Plant Cell Physiol.*, **48**, 391–404.

Zhou,L. *et al.* (2010) Genome-wide identification and analysis of drought-responsive microRNAs in Oryza sativa. *J Exp. Bot.*, **61**, 4157–4168.

Zhou,X. *et al.* (2007) Characterization and identification of microRNA core promoters in four model species, *PLoS Comput Biol*, **3**, e37.

Zhu,Q.H. *et al.* (2008) A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Res.*, **18**, 1456–1465.