Journal of
Data Mining in Genomics & Proteomics

Oliveira et al., J Data Mining Genomics Proteomics 2013, 4:2
DOI: 10.4172/2153-0602.1000130

Research Article Open Access

A Computational Approach for MicroRNA Identification in Plants: Combining Genome-Based Predictions with RNA-Seq Data

Jorge S Oliveira[1#], Nuno D Mendes[1,2#],Victor Carocha[3,4,5], Clara Graça[5,6], Jorge A Paiva[5,6] and Ana T Freitas[1*]

[1]*INESC-ID/IST-Instituto de Engenharia de Sistemas e Computadores/Instituto Superior Técnico, R. Alves Redol 9, 1000 Lisboa, Portugal*
[2]*Network Modelling Group, Instituto Gulbenkian de Ciência, Rua da Quinta Grande, 6 2780-156 Oeiras, Portugal*
[3]*ITQB-Instituto de Tecnologia Química e Biológica, Av. da República, Estação Agronómica Nacional, 2780-157 Oeiras, Portugal*
[4]*LRSV, Laboratoire de Recherche en Sciences Végétales, Université Toulouse III, UPS, CNRS, BP 42617, Auzeville, 31326 Castanet Tolosan, France*
[5]*IICT-Instituto de Investigação Científica e Tropical, Palácio Burnay-Rua da Junqueira, 30, 1349-007 Lisboa, Portugal*
[6]*IBET-Instituto de Biologia Experimental e Tecnológica, Av. da República, Quinta do Marquês, 2781-901 Oeiras, Portugal*
[#]*Authors have contributed equally to the article*

Abstract

MicroRNAs are endogenous molecules that act by silencing targeted messenger RNAs, and which have an important regulatory role in many physiological processes in both plants and animals. Here, we propose a pipeline that makes use of CRAVELA, a single-genome microRNA finding tool originally developed for microRNA discovery in animals, and an NGS data analysis algorithm that provides a novel scoring function to evaluate the expression profile of candidates, taking advantage of the expected relative abundance of RNA fragments originating from the mature sequence, compared to other portions of the microRNA precursor. This approach was tested in *Eucalyptus spp.* for which, despite their economic importance, no microRNAs have been documented. The outcome of our approach was a short list of candidates, including both conserved and non-conserved sequences. Experimental validation showed amplification in 6 out of 8 candidates chosen from the best-scoring non-conserved sequences.

Keywords: MicroRNA; Eucalyptus; *In silico* prediction; CRAVEL; RNA-seq; Plants

Introduction

MicroRNAs (miRNAs) are endogenous ~ 22 nucleotides (nt)-long RNAs that play important regulatory roles in animals and plants by targeting messenger RNAs (mRNAs), for cleavage or translational repression [1]. Although the first miRNAs were identified in 1993 [2], they escaped notice until relatively recently.

MiRNAs comprise one of the most abundant classes of gene regulatory molecules in multicellular organisms, and influence the output of many protein-coding genes [1].

The multitude of small non-coding RNAs (ncRNAs) found in plants (siRNAs, miRNAs, tasiRNAs) complicates the identification of miRNAs. Like animal miRNAs, plant miRNAs (1) are endogenously expressed from one arm of a foldback precursor, (2) are generally conserved in evolution, and (3) come from regions of the genome distinct from previously annotated genes [2].

MiRNA precursors (pre-miRNAs) in plants are much more variable in size than those of animals, ranging from around 60 to a few hundred nucleotides, whereas those in animals are typically 70-nt-long [3], and can have a much more complicated secondary structure, including several bulges. Plant pre-miRNAs are fully processed in the nucleus by protein complexes, and the mature miRNA is then exported to the cytoplasm, where it is incorporated in effector molecular machinery. Plant miRNAs frequently cleave and thus induce immediate degradation of target mRNAs, with target sites, unlike with animal miRNAs, often being located in the coding region, but can also be found in non-coding portions of the gene such as the 3'-UTR, or even the 5'-UTR [2].

Several miRNA discovery methods have been developed specifically for plant genomes [3], but most either rely heavily on a conservation filter, or they are target-centered approaches, being dependent on their ability to confidently identify bonafide miRNA targets. Methods depending on the sieve of evolutionary conservation require a direct comparison with a sequenced genome of a phylogenetically close species; otherwise they are limited to the identification of miRNAs that are extensively conserved across evolution, hampering the discovery of species-specific or genus-specific regulators. More recently, an approach that does not rely on a conservation sieve or the identification of targets–CRAVELA–was proposed [4,5]. Although originally developed for miRNA discovery in animals, the principles it uses are also applicable to plants, as we demonstrate in this paper. The number of candidates CRAVELA identifies is still too numerous to systematically submit for experimental verification; so additional filters need to be introduced.

The CRAVELA processing pipeline is summarized in Figure 1. The primary analysis relies on major characteristics of miRNA precursors: the intrinsic features of pre-miRNAs (such as stability and robustness), the typical secondary structure, based on the structural characteristics of a set of seed precursors, usually chosen from among previously known miRNAs of the given species or taken from homologs of related species. The authors of CRAVELA also propose a third type of analysis referred to as transcriptional potential, since the aim of the tool is to identify genomic sequences with miRNA-like features, but which are not necessarily transcribed. Here, we purport to provide a method to perform this additional analysis. We tested our approach by obtaining precursor candidates from *Eucalyptus spp*. and subsequently using an

***Corresponding author:** Jorge S Oliveira, INESC-ID/IST-Instituto de Engenharia de Sistemas e Computadores/Instituto Superior Técnico, R. Alves Redol 9, 1000 Lisboa, Portugal, E-mail: atf@kdbio.inesc-id.pt

Received May 06, 2013; **Accepted** May 24, 2013; **Published** May 31, 2013

Citation: Oliveira JS, Mendes ND,Carocha V, Graça C, Paiva JA, et al. (2013) A Computational Approach for MicroRNA Identification in Plants: Combining Genome-Based Predictions with RNA-Seq Data. J Data Mining Genomics Proteomics 4: 130. doi:10.4172/2153-0602.1000130

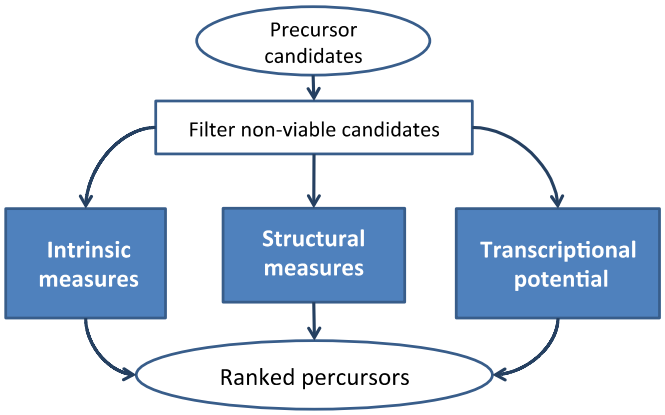

Figure 1: CRAVELA processing pipeline. The left-most module evaluates candidates in terms of stability and robustness criteria, the middle module performs structural analysis and the right-most module includes annotation filtering and data-specific strategies to assess the transcriptional potential.

RNA-Seq dataset produced to study the process of wood formation. A total of 4 out of 5 top-scoring candidates identified using our approach were experimentally validated.

Methods

Overview

The work presented in this paper was developed in the context of the microEGo project, which aims to identify miRNAs involved in the regulation of the *Eucalyptus globulus* tension wood xylogenesis (wood formation). Genomic and transcriptomic data were available for miRNA finding. The global methodological process includes both *in silico* and "bench" approaches. Figure 2 provides an overview of the global methodology. The computational pipeline described in this paper corresponds to the *in silico* tasks represented by green boxes.

The small RNA-Seq (35bp, GA Analyzer, ILLUMINA) dataset was produced with libraries of xylem tissue collected from *E. globulus* specimens, a species very close to *E. grandis*, for which, unlike the former, a sequenced genome is available [6]. Sampling of seasonal wood was made with a population involving a total of 36 trees from three distinct, non-related genotypes (HD161, CN5 and G-70), kindly provided by RAIZ Institute (Portugal). Different samples were collected in each season, and for each condition and each clone, three biological replicates were performed. The reaction wood was obtained from three genotypes (GM2-58, GB3 and MB43), kindly provided by ALTRI Florestal, SA. A total of 36 trees were bent in 4 different time points and harvested along with the nine non-bent trees (controls). A total of three biological replicates were performed for each condition and each clone.

The small RNA-Seq raw data from *E. globulus* was subject to a sequence read clustering and quality assessment, based on a minimum number of reads. At this point, sequences with less than 5 reads or containing documented repeats were eliminated. The RFAM database [7], was also used to filter annotated ncRNAs. Knowing that the mature sequence of a miRNA has a length ranging between 19 and 26 nt, the dataset was trimmed, preserving only clustered reads with a length within the range. This procedure allowed for an aggressive reduction of the number of candidates, keeping only those with minimal guarantees of detectable expression of a putative mature sequence (first three stages of the procedure described in Figure 2, green box 1).

CRAVELA candidate identification

The input of CRAVELA was the genome of *E. grandis*, obtained from Phytozome [6] (version 7.0, annotation 1.0). The primary pipeline produces a ranked list of stable and robust candidates relying on the intrinsic measures [4]. The optimal cut-offs obtained with this procedure for other plants selected approximately the top 10% (data not shown). However, in order to obtain a conservative reduction of the candidate set, we preserved the top 25% (Figure 2, green box 2).

To obtain potential homologs to be used as seeds for the structural analysis [5], a bidirectional best hit (BBH) list was obtained using BLASTn (version 2.2.25), comparing the dataset of selected candidates with miRBase [8] entries (release 17), referring to the documented pre-miRNAs of *Arabidopsis thaliana* (thale cress), *Vitis vinifera* (common grape vine) and *Populus trichocarpa* (black cottonwood), which correspond to widely studied plants. A BBH consists of a pair of sequences (candidate and documented pre-miRNA), which are the best unidirectional BLASTn hit of each other. This procedure was repeated for each organism. In the case of candidates having a BBH with sequences originating from more than one organism, only the best scoring BBH was preserved.

Expression profile

The expression profile of a pre-miRNA candidate refers to the distribution of the number of reads overlapping at each position of the putative precursor. In order to obtain this value for each position, it is necessary to align the sequenced RNAs to a reference genome. Burrows-Wheeler Alignment Tool (BWA) [9] was chosen for this purpose due to its flexibility, for allowing insertions/deletions and exhibiting an overall good performance.

The authors of miRDeep [10], an NGS-based approach to miRNA identification, made the observation that transcriptomic data of a bona fide pre-miRNA, should provide an expression profile having more abundant reads in the portion corresponding to the mature sequence than in the remaining portions of the precursor hairpin, because the latter is a transient structure that is relatively short-lived when compared to the miRNA (Figure 3).

MiRdeep evaluates candidates suggested by the NGS data alone, so it requires several other filters to control for noise. CRAVELA, on the other hand, produces a set of candidates which are already

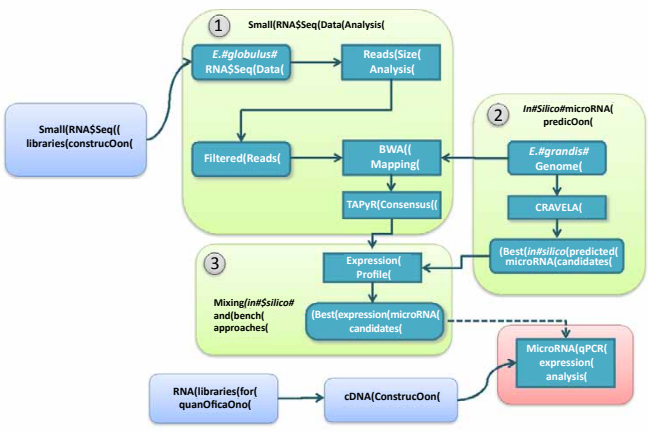

Figure 2: Pipeline developed for *in silico* identification of miRNAs.

filtered following transcription-independent criteria. Inspired by the aforementioned observation, we propose a scoring method, which focuses on the expression signature of the putative precursor. The overall approach is illustrated in Figure 4, it summarizing the data analysis steps performed.

After obtaining the small RNA sequences, as described before, the reads were mapped against the genome of *E. grandis* using BWA. To measure the expression level, we used an option provided by the TAPyR tool [11] (consensus builder), since the relative expression in RNA-Seq data is proportional to the number of complementary DNA (cDNA) fragments that originate from it. This tool option creates consensus sequences and counts the number of reads whose alignment overlaps at each position of the reference genome. This procedure is performed in both directions of the genome, providing the number of reads at each position for each strand. It was possible to use the TAPyR tool to perform this step, since it can read the standard SAM file format produced by BWA (although TAPyR is also a mapping tool, it was designed to work with longer read sizes and is not adapted to very short sequences).

The expression profile scoring is represented in Figure 2 (green box 3), and starts with the identification of the putative mature sequence within the pre-miRNA candidate. As mentioned above, it should correspond to the portion of the precursor with most abundant expression.

Let k(i) denote the number of reads aligned at position i of the genome, p the precursor candidate and p_{start}, p_{stop} its start/stop coordinates, respectively. Algorithm1was used to determine the coordinates of the mature sequence. Intuitively, the precursor is scanned with a sliding window of varying width (19 to 26), which corresponds to the length range of a typical mature miRNA, and determines the combination of position and window width that maximizes the difference between the average number of reads within the window (r_m) and outside the window (r_b).

Algorithm 1 Mature sequence identification

Require: k, p

$R_t \leftarrow \sum_{i=p.\text{start},\ldots,p.\text{stop}} k(i)$
for $w \leftarrow 19$ to 26 **do**
  for $i \leftarrow p.\text{start}$ to $p.\text{stop} - w + 1$ **do**
    $R_m(i,w) \leftarrow \sum_{j=0,\ldots,w-1} k(i+j)$
    $r_m(i,w) \leftarrow \frac{1}{w} R_m(i,w)$
    $r_b(i,w) \leftarrow \frac{1}{p.\text{stop}-p.\text{start}+1-w} (R_t - R_m(i,w))$
return $(i^*, w^*) \leftarrow \arg\max_{i,w} \{r_m(i,w) - r_b(i,w)\}$

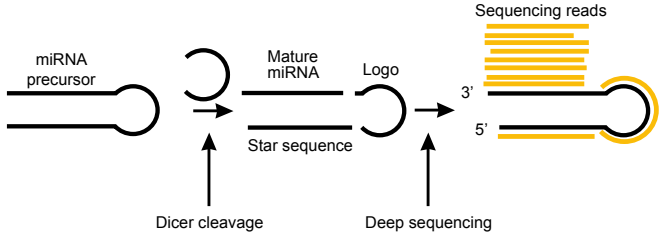

Figure 3: The expected miRNA expression profile - Adapted from [14].

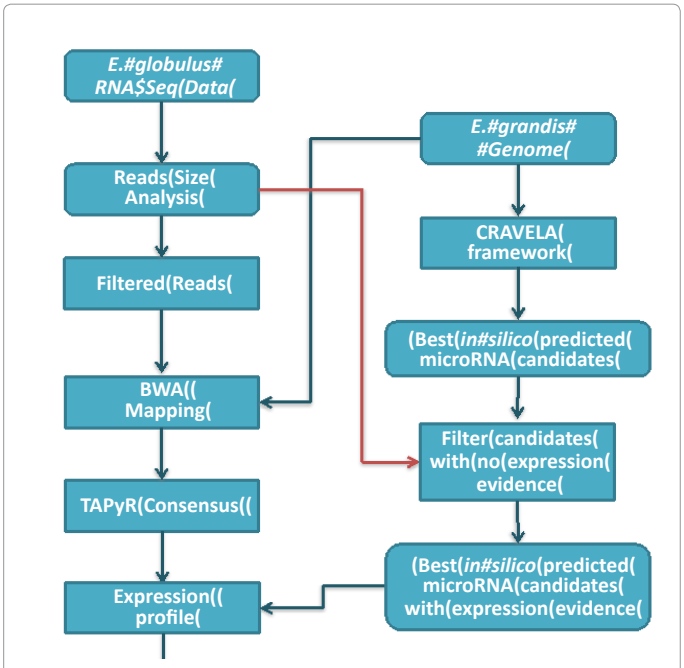

Figure 4: Flowchart with the RNA-Seq data analysis and its integration with the *in silico* analysis.

Relying on the putative mature sequence thus identified and on the secondary structure of the candidate (pre-computed by CRAVELA), it is possible to estimate the identity of the portion of the foldback precursor that is on the opposite stem, termed miRNA*. The interest of identifying the miRNA* is two-fold: (1) occasionally both the miRNA and miRNA* are functional (effectively constituting two alternative mature sequences), and (2) non-functional miRNA* sequences are also expected to be more abundant than the rest of the precursor because they survive longer before degradation.

Having identified both the miRNA and miRNA* sequences, we can obtain a score for the precursor candidate. Let R_m, R_s, and R_t be the total number of reads in the mature, miRNA*, and the whole candidate, respectively. The precursor score, σ, is thus

$$\sigma(\rho) = \frac{Rm + Rs}{Rt} \quad (1)$$

i.e. the proportion of read counts in the precursor located specifically in either the miRNA or miRNA* putative sequences, and consequently, $0 \leq \sigma \leq 1$.

Biological validation

A miRNA-specific RT-qPCR protocol consisting in a stem-loop RT, followed by an end-point qPCR was performed for the validation of mature sequences [12].The validation was carried out using Xylem, Ovaries and Seedling tissue, harvested from non-bent trees. Random primers were used for the construction of the cDNA libraries, and the primers for qPCR were designed using Prime3Plus web application [13]. All the miRNA primers that we used in this validation are compiled in Table 1.

Results

The CRAVELA framework identifies an initial set of over 4 million candidates, which are subsequently drastically reduced by

the application of scoring and filtering procedures. After ranking the precursors with the intrinsic measures and conservatively keeping the top 25%, we produced a reduced list of ~ 1 million candidates. This list is compared using BLASTn with the precursor sequences on miRBase for three related plant species, in order to identify a small set of potential homologs to seed the structural analysis step. The BLASTn procedure yielded approximately 600 matches; the unique bidirectional best hits were computed, as described in the CRAVELA candidate identification section. The found homologs and their respective organism and sequence are compiled in Table 2.

A list of 21 unique homologs, supported by very low E-values (maximum of 10^{-7}), was obtained and used in the structural analysis, eliciting the determination of a set of 380,000 candidates, which were than further reduced using the procedure described in the RNA-Seq primary data analysis section. The primary RNA-Seq data

miR-156	miRNA Sequence	UGACAGAAGAGAGUGAGCAC
	Stem-loop Primer	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACGTGCTC
	qPCR Forward Primer	GCGGCGGTGACAGAAGAGAGT
Universal	qPCR Reverse Primer	GTGCAGGGTCCGAGGT
ath miR-157a	miRNA Sequence	UUGACAGAAGAUAGAGAGCAC
	Stem-loop Primer	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACGTGCTC
	qPCR Forward Primer	GCGGCGGTTGACAGAAGATAGA
ptc miR-319a	miRNA Sequence	UUGGACUGAAGGGAGCUCCC
	Stem-loop Primer	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACGGGAGC
	qPCR Forward Primer	CGGCGGTTGGACTGAAGGGA
novel miR-8851488	miRNA Sequence	GUAUUGGAGUGAAGGGAGCUCCUC
	Stem-loop Primer	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACGAGGAG
	qPCR Forward Primer	GTCAGGTATTGGAGTGAAGGGAG
novel miR-6446782	miRNA Sequence	UCUCGGACCAGGCUUCAUUCCCC
	Stem-loop Primer	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACGGGGAA
	qPCR Forward Primer	GATTCTCTCGGACCAGGCTTCA
novel miR-7602121	miRNA Sequence	CUAACUCGGGAGGUUAGGAGGUCAG
	Stem-loop Primer	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACCCAGTC
	qPCR Forward Primer	GATGATACTAACTCGGGAGGTTAGGA
novel miR-9336066	miRNA Sequence	UCACGAGAGAUAGAAGACAGUUGU
	Stem-loop Primer	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACACAACT
	qPCR Forward Primer	GACGGTCACGAGAGATAGAAGAC
novel miR-8853075	miRNA Sequence	CUAGAAGAACUUGGGGGAGUGCGAA
	Stem-loop Primer	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACTTCGCA
	qPCR Forward Primer	ATCAGCTAGAAGAACTTGGGGGAG
pre-mir156	qPCR Forward Primer	GAGTGGTGAGGAATTGATGG
	qPCR Reverse Primer	GGGGGTGACGGATAGAGAGT
pre-mir157	qPCR Forward Primer	TGATGAGATACAATTCGGAGCA
	qPCR Reverse Primer	AAGGCTAGAGAGCACAAAGGA
pre-mir319	qPCR Forward Primer	GCCGACTCATTCATCCAAAT
	qPCR Reverse Primer	GGAGCTCCCTTCAGTCCAAG
pre-mir-8851488	qPCR Forward Primer	GGGCTCTCAACACTCCATGT
	qPCR Reverse Primer	TCATCTCCAAGTCCATGCAA
pre-mir-6446782	qPCR Forward Primer	AGTTGAGGGGGATGCTGTCT
	qPCR Reverse Primer	CAAGTTGAGGGGAATGAAGC
pre-mir-7602121	qPCR Forward Primer	TCCCTTTTGCTCGATTATGG
	qPCR Reverse Primer	GATTGAGCCCTCCAATCCTC
pre-mir-9336066	qPCR Forward Primer	TGAATGGGGTGTTGACAGAA
	qPCR Reverse Primer	ACAGAGGAGGGAGAGCACAA
pre-mir-8853075	qPCR Forward Primer	CAAGGGCTTCTGGCCTATTA
	qPCR Reverse Primer	AATTAGCAAGGGCACGGTTT
U6	qPCR Forward Primer	CTCGCTTCGGCAGCACA
	qPCR Reverse Primer	AACGCTTCACGAATTTGCGT

Table 1: MiRNA conserved and novel candidates and respective primers designed for the RT-qPCR assays.

analysis, which aims at identifying candidates with strict guarantees of meaningful expression, produced a list of 3,300 candidates, which were then scored using the precursor scoring strategy described in the expression profile section. A total of 300 high-scoring candidates (score>0.85) were thus obtained, from which a few were included in a set submitted for experimental validation.

A total of 8 candidates were selected from the high-scoring list and the potential homologs list, relying on different sources and with different justifications. The information regarding the selection of candidates is summarized in Table 3.

The RT-qPCR assays for the pre- and mature miRNAs showed amplification in almost all the tested candidates as seen in Table 4. The miR-156 candidate did not show any selective amplification probably due to the high sequence similarity to the miR-157 candidate, or due to limitations of the technology.

The overall success was lower, while attempting precursor validation (4 out of 7), probably because of its high cellular turnover and therefore, its lower abundance. Since the goal of this paper was to present an approach that puts together *in silico* and bench procedures to analyze high-throughput data, we decided not to include a discussion of the biological significance of the obtained results. However, it is important to show that the computational approach lead to results that guided the bench experiments with meaningful biological findings.

Conclusions and Future Work

Recent advances in high-throughput sequencing technology allow the production of fast and abundant transcription data. Computer science, mathematics and statistics are essential fields for the handling of this output, and the integration of these data with genomic knowledge in an effort to unravel gene function and regulatory interactions.

The use of a mixed approach combining *in silico* pre-miRNA predictions and RNA-Seq transcriptional expression profiling elicited the production of a first catalogue of 3,300 putative conserved and novel miRNA candidates. The proposed scoring mechanism ranks the candidate list based on the similarity of the experimental data with a typical miRNA deep-sequencing expression profile. Therefore, it allows to refine the candidate list by discarding those not behaving like miRNAs, and identifying a restricted set of putative precursors that might be used in further experimental research. Using an RT-qPCR protocol specific for miRNAs, it was possible to experimentally verify, with success, the expression of both predicted homologs and a

Identifier	Organism	Mature sequence
miR156f	*A. thaliana*	UGACAGAAGAGAGUGAGCAC
miR157a	*A. thaliana*	UUGACAGAAGAUAGAGAGCAC
miR157b	*A. thaliana*	UUGACAGAAGAUAGAGAGCAC
miR156b	*P. trichocarpa*	UGACAGAAGAGAGUGAGCAC
miR160c	*P. trichocarpa*	UGCCUGGCUCCCUGUAUGCCA
miR171c	*P. trichocarpa*	AGAUUGAGCCGCGCCAAUAUC
miR171d	*P. trichocarpa*	AGAUUGAGCCGCGCCAAUAUC
miR171g	*P. trichocarpa*	UGAUUGAGCCGUGCCAAUAUC
miR171l	*P. trichocarpa*	CGAGCCGAAUCAAUAUCACU
miR171n	*P. trichocarpa*	CGAGCCGAAUCAAUAUCACU
miR319a	*P. trichocarpa*	UUGGACUGAAGGGAGCUCCC
miR319g	*P. trichocarpa*	UUGGACUGAAGGGAGCUCCU
miR156c	*V. vinifera*	UGACAGAAGAGAGUGAGCAC
miR156g	*V. vinifera*	UUGACAGAAGAUAGAGAGCAC
miR156i	*V. vinifera*	UUGACAGAAGAUAGAGAGCAC
miR171a	*V. vinifera*	UGAUUGAGCCGUGCCAAUAUC
miR171f	*V. vinifera*	UUGAGCCGCGCCAAUAUCACU
miR171i	*V. vinifera*	UGAUUGAGCCGUGCCAAUAUC
miR319c	*V. vinifera*	UUGGACUGAAGGGAGCUCCCU
miR399h	*V. vinifera*	UGCCAAAGGAGAAUUGCCCUG
miR535c	*V. vinifera*	UGACAACGAGAGAGAGCACGC

Table 2: List of homologs found through the BLASTn between the *E. grandis*genome and the miRBase dataset for three plant species. Although the entire precursor was used as a homolog, we have chosen to present only the mature sequence due to size restrictions.

Code/ Identifier	Source	Gene Family	Putative Function/Target	Selection Criterion
miR-156	*P. trichocarpa* homolog	MIR156	Floral dev. [21]	Widely described, Test selectivity[a]
miR-157	*A. thaliana* homolog	MIR156	Floral dev. [21]	Widely described, Test selectivity[a]
miR-319	*A. thaliana* homolog	MIR159	Leaves/fruits morphology [22]	Widely described
miR-CH1	Pipeline candidate	MIR159	Targets transcription factors [23]	Interesting function
miR-CH2	Pipeline candidate	MIR166	Cambium/xylem diff.[24]	Interesting function
miR-LB1	Pipeline candidate	Unknown	Unknown	High scored in expression profile
miR-LB2	Pipeline candidate	Unknown	Unknown	High scored in expression profile
miR-C09	Northern validated	MIR477	rRNA processing [25]	Northern validation, Interesting function

[a]Candidates miR-156 and miR-157 have very similar sequences. They were both included to test the RT-qPCR selectivity power.

Table 3: Biological validation, putative function of selected miRNAs and selection criterion

Candidate	Precursor amplification	Mature amplification
miR-156	n/a	No
miR-157	No	Yes
miR-319	Yes	Yes
miR-CH1	Yes	Yes
miR-CH2	No	Yes
miR-LB1	No	Yes
miR-LB2	Yes	Yes
miR-C09	Yes	No

Table 4: RT-qPCR amplification results for the tested candidates.

few high-scoring non-conserved candidates produced by the miRNA discovery pipeline. The proposed computational pipeline, relying on CRAVELA for obtaining an initial list of candidates, presented results with encouraging accuracy in the top-scoring candidates selected for experimental verification. It would thus be of interest to perform the same data analysis on other economically important species such as *V. vinifera*, *P. trichocarpa* (a model tree), and other plant models, such as *A. thaliana* and *Medicago truncatula*.

A critical step in the functional annotation of novel miRNAs is the identification of their targets. Despite the fact that plant miRNAs tend to be fully complementary to their target sites, which facilitates their computational enumeration, the short length of mature sequences leads to an excessive number of spurious matches, which result in a large number of false positive results. To complement the pipeline presented in this paper with a means to identify potential targets, notably mRNAs being putatively targeted by more than one candidate, could provide a network-based method to not only improve the accuracy of target prediction, but also to filter candidates with no confident targets.

As mentioned previously, the methodology described in this paper was applied using *E. grandis* genome assembly, while combining it with *E. globulus* RNA-Seq data. Sequence similarity between the *E. grandis* and *E. globulus* genomes proved to be sufficient to support transcriptional analysis on the latter, both in terms of the outcome of miRNA predictions, but also for the alignment of RNA-Seq reads with the reference genome. The use of annotation data would provide extra information, eliciting the exclusion of pre-miRNA candidates overlapping annotated genes, repeat sequences or other ncRNAs. Despite the evidence supporting high levels of conservation in many gene families between the genomes of *E. grandis* and *E. globulus*, it would be preferable to use the *E. globulus* genome. The sequencing and annotation of *E. globulus* would be an invaluable contribution to the research efforts seeking the identification and characterization of regulators of wood formation in this species, without which we cannot identify miRNAs specific to *E. globulus*.

This article was originally published in a special issue, **Bioinformatics for Highthroughput Sequencing** handled by Editor: Dr. Heinz Ulli Weier, Lawrence Berkeley National Laboratory, USA

Acknowledgments

This work was supported by national funds through FCT, under project (PTDC/AGR-GPL/098179/2008, P-KBBE/AGR- GPL/0001/2010, PTDC/EIA-EIA/122534/2010 and PEst-OE/EEI/LA0021/2011). Jorge S Oliveira and Jorge A Paiva acknowledge the microEGo project BI fellowship and the ProgramaCiência 2008, funded by POPH(QREN).We acknowledge Portugal ALTRI FLORESTAL SA and RAIZ, for providing the plant material. Finally, the authors acknowledgethe EUCAGEN consortium, A. Myburg, D. Grattapagliaand J. Tuskanand the US-Department of Energy, for making available the *E. grandis* genome sequence.

References

1. Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism and function. Cell 116: 281-297.
2. Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell 75: 843-854.
3. Mendes ND, FreitasAT, Sagot MF (2009) Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res 37: 2419-2433.
4. Mendes ND, Freitas AT, Vasconcelos AT, Sagot MF (2010) Combination of measures distinguishes pre-miRNAs from other stem-loops in the genome of the newly sequenced *Anopheles darlingi*. BMC Genomics 11: 529.
5. Mendes ND, Heyne S, Freitas AT, Sagot MF, Backofen R (2012) Navigating the unexplored seascape of pre-miRNA candidates in single-genome approaches. Bioinformatics 28: 3034-3041.
6. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, et al. (2012) Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res 40: D1178-D1186.
7. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Res 31: 439-441.
8. Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Res 32: D109-D111.
9. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
10. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, et al. (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 26: 407-415.
11. Fernandes F, da Fonseca PG, Russo LM, Oliveira AL, Freitas AT (2011) Efficient alignment of pyrosequencing reads for re-sequencing applications. BMC Bioinformatics 12: 163.
12. Varkonyi-Gasic E, Wu R, Wood M, Walton EF, Hellens RP (2007) Protocol: a highly sensitive RT-PCR method for detection and quantification of microRNAs. Plant Methods 3: 12.
13. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, et al. (2007) Primer3plus, an enhanced web interface to primer3. Nucleic Acids Res 35: W71-W74.
14. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, et al. (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 26: 407-415.