Supplement

to

NOVOMIR:

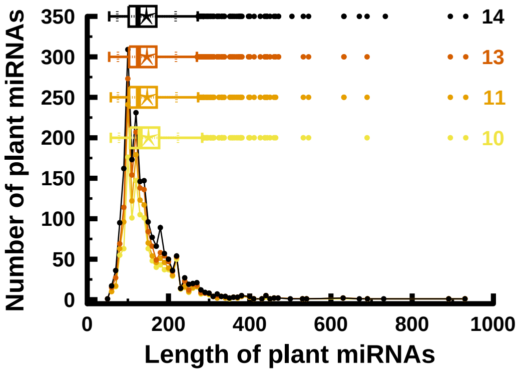
de novo prediction of microRNA-coding regions in a single plant genome

Jan-Hendrik Teune and Gerhard Steger*

Heinrich-Heine-Universität Düsseldorf, Institut für Physikalische Biologie, Universitätsstr. 1,
D-40225 Düsseldorf, Germany

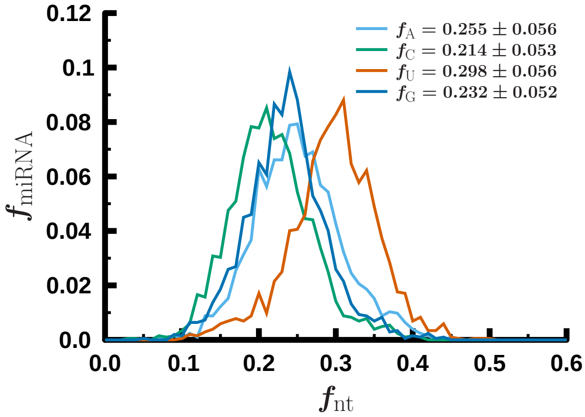
∗ Corresponding author; Tel.: +211 81 14597; Fax: +211 81 15167

E-mail addresses: teune@biophys.uni-duesseldorf.de; steger@biophys.uni-duesseldorf.de



MIRBASE	length/nt			
version	mean	Q1	median	Q3
	plant pre-miRNAs			
10.0	150.8±72.6	105	130	177
11.0	147.6±72.1	103	126	171
13.0	147.0±71.5	103	126	170
14.0	145.7±71.9	102	125	170
	non-plant pre-miRNAs			
10.0	87.7±14.0	78	86	97
11.0	87.2±15.6	77	86	97
13.0	89.0±15.8	78	87	97
14.0	87.2±15.9	76	86	97

Figure S1: Length distribution of pre-miRNA in different versions of MIRBASE. Left: Data for plant pre-miRNAs in MIRBASE versions 10, 11, 13, and 14 are given in yellow, orange, red, and black, respectively. Mean and standard deviation are marked by a star and dotted lines. Median (thick vertical bar), quartiles, and outliers are depicted as box plots. Right: Values in tabular form.



MIRBASE version	f_{nt} A	C	U	G
	plant pre-miRNAs			
10.0	0.249±0.056	0.221±0.051	0.291±0.057	0.239±0.051
11.0	0.249±0.054	0.220±0.051	0.294±0.055	0.237±0.049
13.0	0.252±0.056	0.217±0.053	0.296±0.056	0.234±0.051
14.0	0.255±0.056	0.214±0.053	0.298±0.056	0.232±0.052
	mature plant miRNAs			
10.0	0.242±0.098	0.234±0.106	0.260±0.100	0.265±0.098
11.0	0.242±0.097	0.237±0.104	0.257±0.099	0.264±0.097
13.0	0.245±0.099	0.231±0.105	0.260±0.102	0.264±0.099
14.0	0.245±0.099	0.233±0.107	0.262±0.102	0.260±0.101

Figure S2: Nucleotide composition of plant pre-miRNAs and mature miRNAs in different versions of MIRBASE. Top: Data for pre-miRNAs in MIRBASE version 14. Bottom: Data in tabulated form for MIRBASE versions 10, 11, 13, and 14.

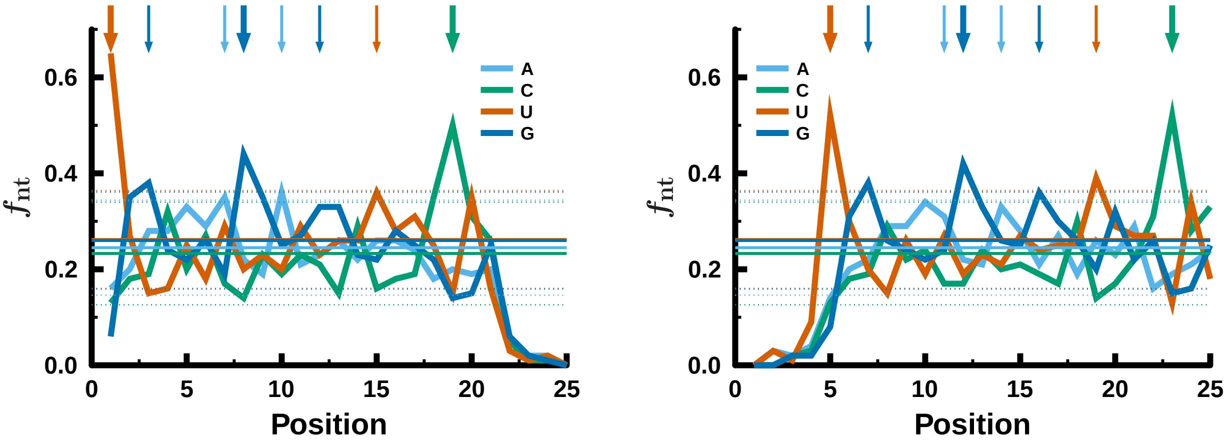


Figure S3: Nucleotide distribution in mature plant miRNAs from MIRBASE version 14. The mean length is 22.2 nt; the length of the longest sequence is 25 nt. The nucleotide distribution was determined either starting from the 5' end (left) or from the 3' end (right) of all mature miRNAs. The horizontal lines mark the mean nucleotide content of miRNAs (solid) and its standard deviation (dotted). The arrows mark positions with frequencies of a certain nucleotide above mean plus standard deviation.

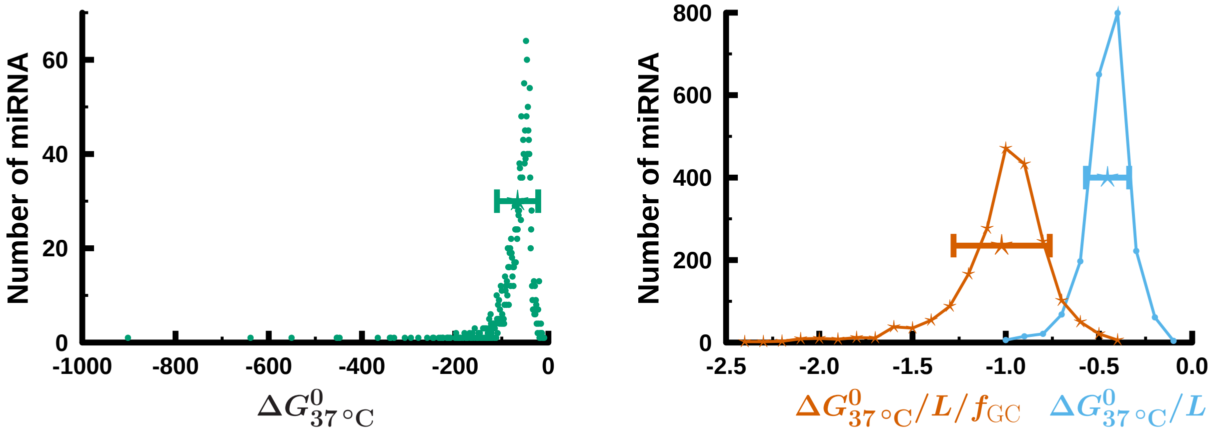


Figure S4: Minimum of free energy (in kcal/mol) of secondary structures of plant pre-miRNA in MIRBASE version 14. Calculations were done by RNAFOLD [1] using default parameters. Energies were binned into 1 kcal/mol (left) or in 0.1 kcal/mol (right) bins. In the right plot, energy values are normalized to the length of pre-miRNA ($\Delta G/L$) and to GC-content of the pre-miRNA sequences ($\Delta G/L/f_\mathrm{GC}$), respectively.

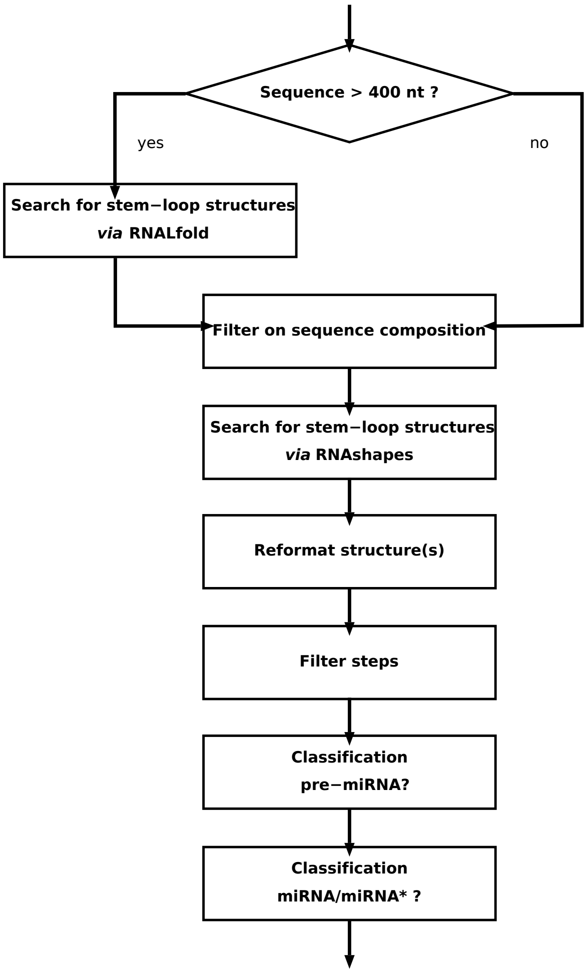

Sequence > 400 nt ?
yes
no
Search for stem−loop structures *via* RNALfold
Filter on sequence composition
Search for stem−loop structures *via* RNAshapes
Reformat structure(s)
Filter steps
Classification pre−miRNA?
Classification miRNA/miRNA* ?

Figure S5: Workflow in NOVOMIR. RNALFOLD [2] is used to search for sub-sequences with local stem-loop structures if the sequence is longer than 500 nt. Each sub-sequence or the original sequence (shorter than 500 nt) is rejected if it contains a strongly biased nucleotide composition. RNASHAPES [3] is used to predict structures for the sequence up to a certain energy threshold. Predicted structures are reformatted into an alignment-like format (see Fig. 1). If a structure satisfies several filter criteria, NOVOMIR calculates the probability that the sequence (and its structure) might be a pre-miRNA. If so, NOVOMIR predicts the localization of the miRNA/miRNA* complex in the presumed pre-miRNA.

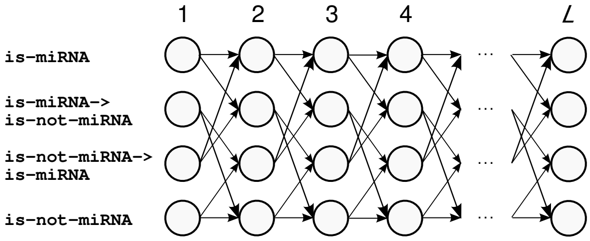


Figure S6: Hidden-Markov model used for classification.

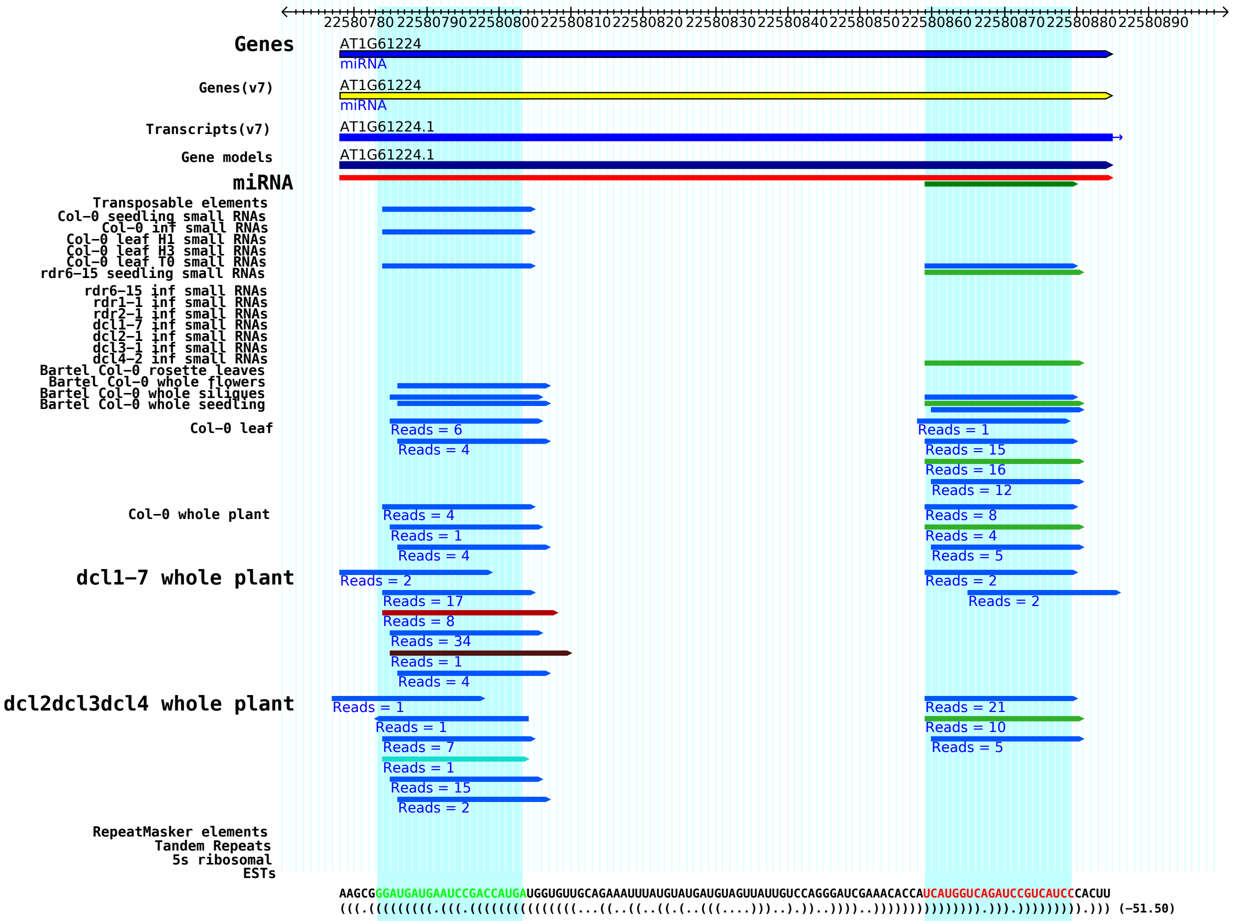
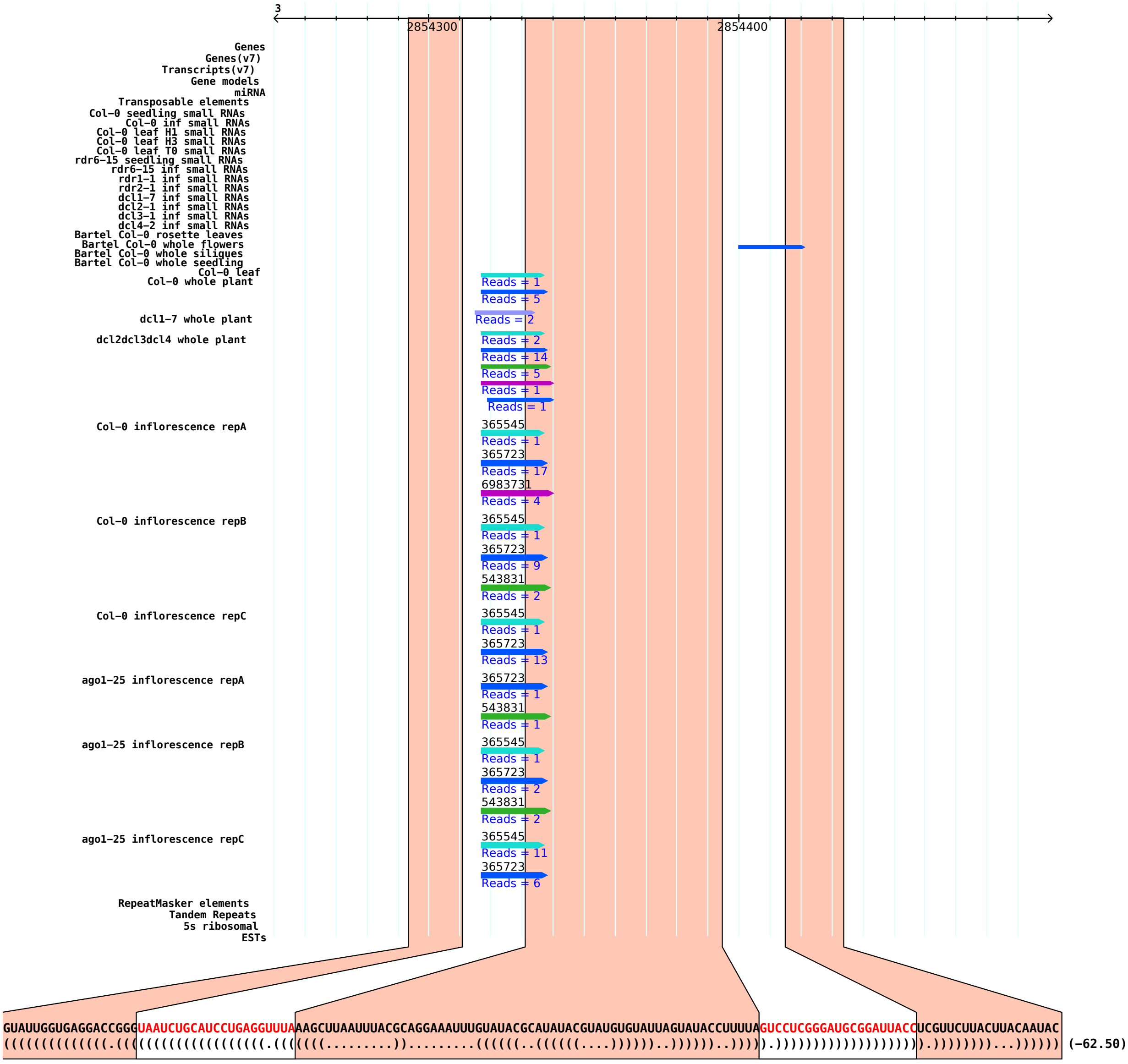


Figure S7: Expression pattern of ath-MIR842a viewed in the ASRP browser. This screen shot from the ASRP browser at `http://asrp.cgrb.oregonstate.edu/cgi-bin/gbrowse/thaliana/` shows the genomic sequence region from nt 22,580,770 to nt 22,580,900 of *A. thaliana* chromosome 1, which includes the MIR842a gene (AT1G61224). At the bottom the secondary structure of ath-MIR842a is added as predicted by RNAFOLD; the mature miRNA sequence is marked in red and its complementary region in green. Note the presence of miRNA and miRNA* sequences especially in the plants that are defective in Dicer-like proteins (dcl1-7, dcl2dcl3dcl4). The colors of the expressed small RNAs denote their lengths: **19**, **20**, **21**, **22**, **23**, **24**, and **25** nt.



```
GUAUUGGUGAGGACCGGGUAAUCUGCAUCCUGAGGUUUAAAGCUUAAUUUACGCAGGAAAUUUGUAUACGCAUAUACGUAUGUGUAUUAGUAUACCUUUUAGUCCUCGGGAUGCGGAUUACCUCGUUCUUACUUACAAUAC
((((((((((((.(((((((((((((((((.((((((.........))).........((((((..((((((....))))))..))))))..)))))).)))))))))))))))))))).))))))))...)))))) (-62.50)
```

Figure S8: Expression pattern of a pre-miRNA candidate viewed in the ASRP browser. This screen shot shows the genomic sequence region from nt 2,854,250 to nt 2,854,500 of *A. thaliana* chromosome 3. The reddish bars map the pre-miRNA to the genome; the intermediate white bars mark the predicted regions of the mature miRNA and the miRNA*. For further details see Fig. S7.

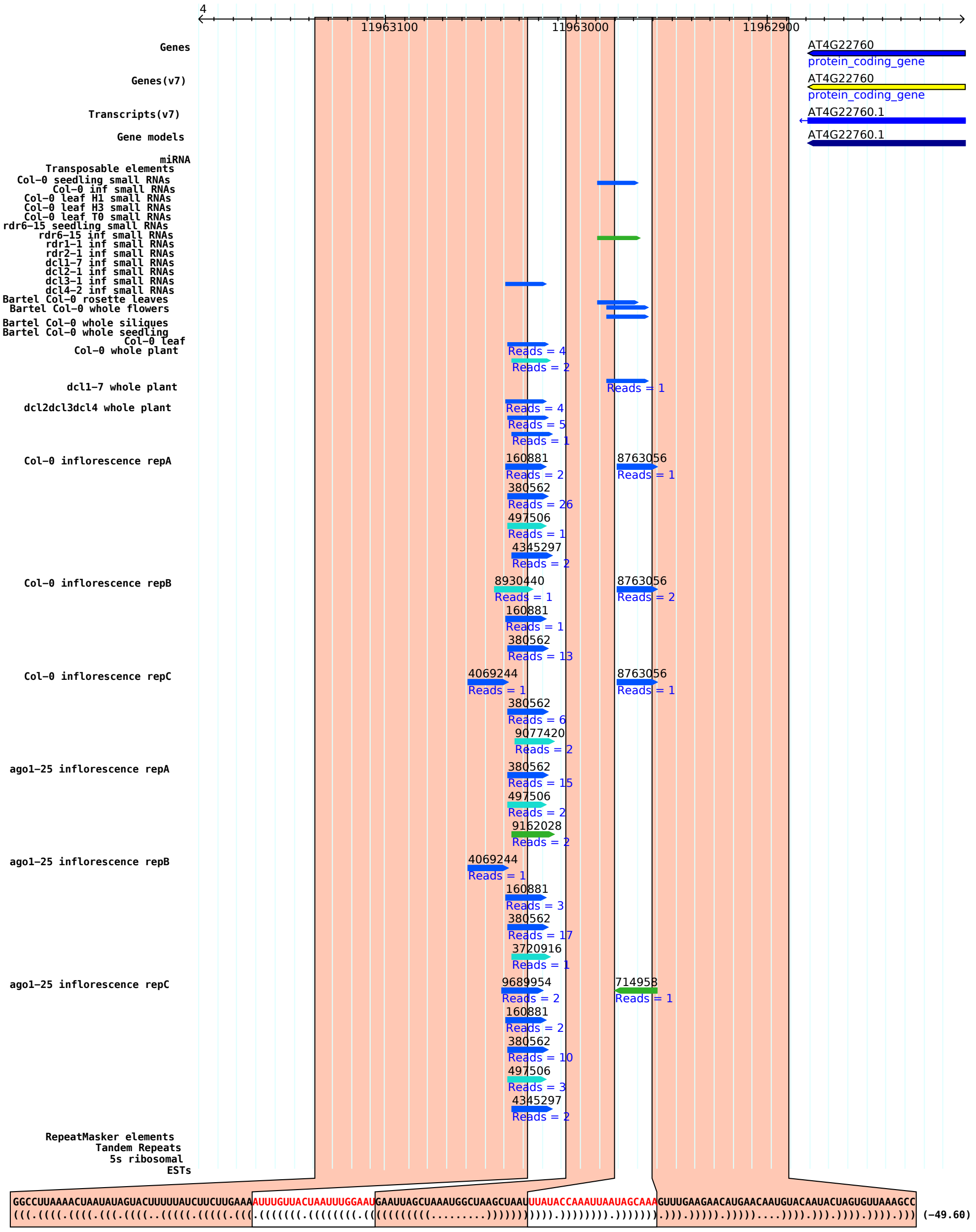


Figure S9: Expression pattern of a pre-miRNA candidate viewed in the ASRP browser. This screen shot shows the genomic sequence region from nt 11,963,200 to nt 11,962,800 of *A. thaliana* chromosome 4. For further details see Fig. S7 and Fig. S8.

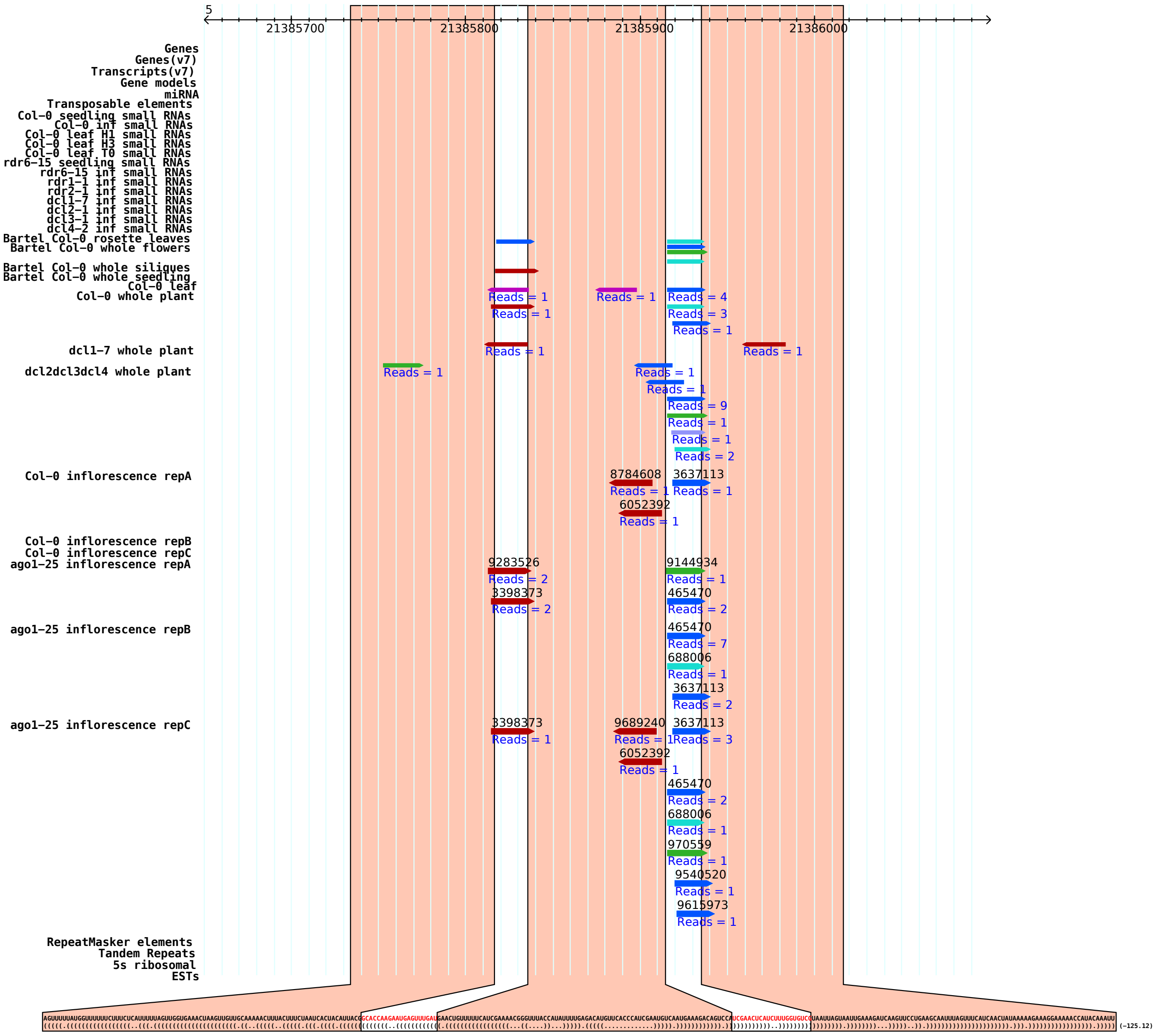


Figure S10: Expression pattern of a pre-miRNA candidate viewed in the ASRP browser. This screen shot shows the genomic sequence region from nt 21,385,650 to nt 21,386,100 of *A. thaliana* chromosome 5. For further details see Fig. S7 and Fig. S8.

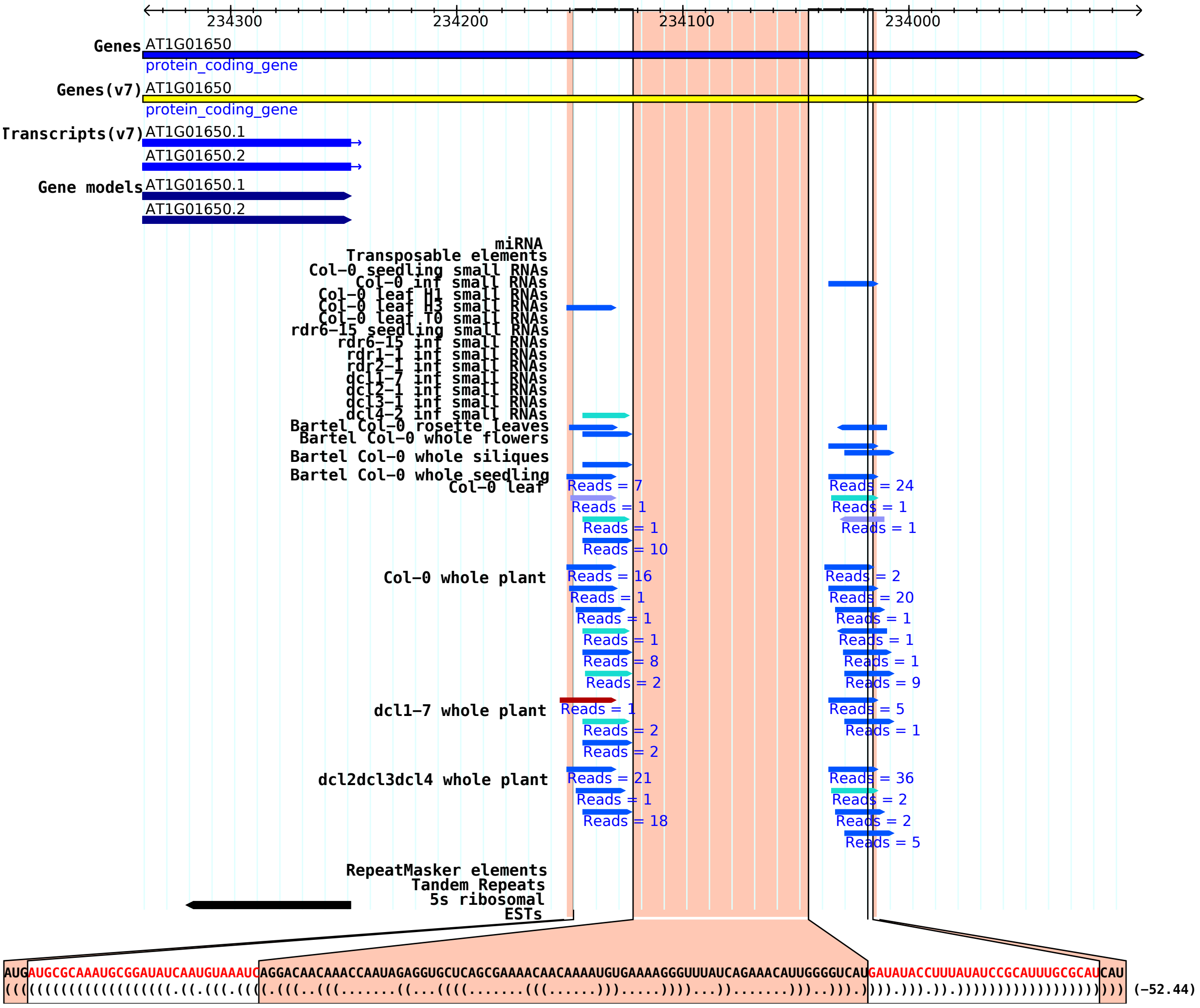


Figure S11: Expression pattern of a pre-miRNA candidate viewed in the ASRP browser. This screen shot shows the genomic sequence region from nt 234,340 to nt 233,900 of *A. thaliana* chromosome 1; an intron of AT1G01650, an aspartic-type endopeptidase/peptidase, extends from nt 233618–234250. For further details see Fig. S7 and Fig. S8.

Table S1: Sensitivity of NOVOMIR. Comparison of NOVOMIR's sensitivity for detecting plant pre-miRNAs from different versions of MIRBASE. The column "MIRBASE 14 - MIRBASE 10" shows values for sequences from MIRBASE 14 which are not present in MIRBASE 10. Note that NOVOMIR's thresholds and probabilities were learned only from *A. thaliana* sequences in MIRBASE version 10.

	MIRBASE 10				MIRBASE 14				MIRBASE 14 - MIRBASE 10			
Species	# sequences	TP	FN	Sensitivity	# sequences	TP	FN	Sensitivity	# sequences	TP	FN	Sensitivity
plant	1247	987	260	0.79	2030	1608	422	0.79	788	627	161	0.80
Arabidopsis thaliana	184	154	30	0.84	190	158	32	0.83	6	4	2	0.67
Aquilegia coerulea	0	–	–	–	45	36	9	0.80				
Brachypodium distachyon	0	–	–	–	19	5	14	0.26				
Brassica napus	4	3	1	0.75	45	40	5	0.89	42	37	5	0.88
Brassica oleracea	0	–	–	–	5	5	0	1.00				
Brassica rapa	0	–	–	–	17	16	1	0.94				
Carica papaya	0	–	–	–	1	0	1	0.00				
Chlamydomonas reinhardtii	43	37	6	0.86	44	38	6	0.86				
Gossypium herbecium	0	–	–	–	1	1	0	1.00				
Gossypium hirsutum	0	–	–	–	13	11	2	0.85				
Gossypium rammindii	0	–	–	–	2	2	0	1.00				
Glycine max	22	19	3	0.86	85	67	18	0.79	65	50	15	0.77
Lotus japonicus	0	–	–	–	2	2	0	1.00				
Malus domestica	0	–	–	–	1	1	0	1.00				
Medicago truncatula	30	27	3	0.90	108	100	8	0.93	78	73	5	0.94
Oryza sativa	243	199	44	0.82	414	338	76	0.82	173	142	31	0.82
Populus euphratica	0	–	–	–	8	2	6	0.25				
Physcomitrella patens	220	166	54	0.76	230	175	55	0.76	10	9	1	0.90
Pinus taeda	27	21	6	0.78	37	23	14	0.62	10	2	8	0.20
Populus trichocarpa	215	154	61	0.72	234	169	65	0.72	19	15	4	0.79
Phaseolus vulgaris	0	–	–	–	8	6	2	0.75				
Sorghum bicolor	66	52	14	0.79	140	106	34	0.76	74	55	19	0.74
Solanum lycopersicum	0	–	–	–	30	27	3	0.90				
Selaginella moellendorffii	57	53	4	0.93	57	53	4	0.93				
Saccharum officinarum	16	8	8	0.50	16	8	8	0.50				
Triticum aestivum	28	20	8	0.71	31	22	9	0.71				
Vigna unguiculata	0	–	–	–	2	2	0	1.00				
Vitis vinifera	0	–	–	–	140	110	30	0.79				
Zea mays	92	74	18	0.80	105	85	20	0.81	13	11	2	0.85

References

[1] Hofacker, I. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

[2] Hofacker, I., Priwitzer, B., and Stadler, P. (2004). Prediciton of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.

[3] Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. (2006). RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.