

Gökhan Karakulah¹

Discovery and Annotation of Plant Endogenous Target Mimicry Sequences from Public Transcriptome Libraries: A Case Study of *Prunus persica*

¹ İzmir International Biomedicine and Genome Institute (İBG-İzmir), Dokuz Eylül University, 35340, İnciraltı, İzmir, Turkey, Tel.: +90 232 412 6537, E-mail: gokhan.karakulah@deu.edu.tr

Abstract:

Novel transcript discovery through RNA sequencing has substantially improved our understanding of the transcriptome dynamics of biological systems. Endogenous target mimicry (eTM) transcripts, a novel class of regulatory molecules, bind to their target microRNAs (miRNAs) by base pairing and block their biological activity. The objective of this study was to provide a computational analysis framework for the prediction of putative eTM sequences in plants, and as an example, to discover previously un-annotated eTMs in *Prunus persica* (peach) transcriptome. Therefore, two public peach transcriptome libraries downloaded from Sequence Read Archive (SRA) and a previously published set of long non-coding RNAs (lncRNAs) were investigated with multi-step analysis pipeline, and 44 putative eTMs were found. Additionally, an eTM-miRNA-mRNA regulatory network module associated with peach fruit organ development was built via integration of the miRNA target information and predicted eTM-miRNA interactions. My findings suggest that one of the most widely expressed miRNA families among diverse plant species, miR156, might be potentially sponged by seven putative eTMs. Besides, the study indicates eTMs potentially play roles in the regulation of development processes in peach fruit via targeting specific miRNAs. In conclusion, by following the step-by-step instructions provided in this study, novel eTMs can be identified and annotated effectively in public plant transcriptome libraries.

Keywords: Endogenous target mimicry, microRNA, non-coding RNA, novel transcript discovery, post-transcriptional regulation

DOI: 10.1515/jib-2017-0009


Received: March 8, 2017; **Revised:** March 30, 2017; **Accepted:** April 12, 2017

1 Introduction

RNA sequencing is a powerful tool to identify distinct types of previously un-annotated transcripts, including coding and non-coding RNAs, and new splice isoforms of known genes, in addition to measurement of gene and/or transcript abundances precisely [1], [2], [3]. *De novo* transcriptome reconstruction is now becoming routine and substantially improving the transcriptome annotation of animal and plant species [1], [4]. Moreover, the identification of such novel genetic sequences helps us to better understand complex gene regulatory networks that underlie biological processes, such as development and disease [5], [6], [7]. Therefore, discovery of novel transcripts has recently begun to receive growing attention by the scientific community, and several computational methodologies have been developed so far to annotate types of transcripts [2], [8], [9], [10], [11], [12], [13].

MicroRNA (miRNA) as a class of non-coding regulatory RNAs is involved in the regulation of gene expression in eukaryotes [14]. It is known that the expression of ~60 % of the all coding transcripts in mammals are under the control of miRNA-based regulation [15]. To date a total of 28,265 miRNAs have been deposited in a public database called miRBase (<http://www.mirbase.org>) [16]. Post-transcriptional suppression activity of miRNAs can also be controlled via endogenous target mimicry (eTM) [17], [18]. ETM molecules, also called miRNA sponges or competing endogenous RNAs (ceRNAs), a type of long non-coding RNA (lncRNA), specifically block the targeting activity of miRNAs to decoy the miRNA-based target transcript suppression [19]. In target mimicry, eTM transcript binds to target miRNA with perfect base pairing from the 2nd to 8th nucleotides,

Gökhan Karakulah is the corresponding author.

 ©2017, Gökhan Karakulah, published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License.

and with a three-nucleotide bulge from nucleotides 9–11 at the 5' end of target miRNA sequence. Thus, the eTM-mRNA pairing inhibits miRNA activity and prevents the binding of the miRNA to its target mRNAs, which results in increased levels of target mRNA expression.

A number of plant eTM sequences have been computationally predicted so far [20], [21], [22], [23], [24], [25], and some of them have been functionally characterized [18], [21], [26], [27]. In a recent report by [27], several eTM transcripts were computationally identified in *Glycine max* (soybean) through RNA sequencing, and it was shown that nine miRNAs involved in lipid metabolism were potentially regulated by previously un-annotated eTMs. Besides, the expression level of miRX27 involved in nicotine biosynthesis was sponged by up-regulation of an eTM (nta)-eTMX27 in *Nicotiana tabacum* (tobacco) [26]. Recently, PeTMbase [24] and PceRBase [25] online tools have been established to facilitate the study of putative eTM-miRNA interactions in plants. PeTMbase permits searching and browsing of computationally identified ~2700 eTM transcripts in 11 plant species. The database annotates eTM transcripts through previously described binding rules by [21], and it allows retrieving eTM sequences and their genomic features by given target miRNA and species name. PceRBase is a comprehensive catalogue of target mimicry sequences that comprises more than 160,000 target-mimic pairs (as of March 2017) determined through Tapir tool [28] from 26 plant species. One of the unique features of PceRBase is to provide its users with detailed information regarding putative eTM-miRNA pairs, including their expression levels and associated Gene Ontology (GO) terms. MiRSponge [29] is another source of target mimicry molecules experimentally validated in diverse organisms. However, it has a very limited number of eTM data for only *Arabidopsis thaliana*.

Given the regulatory roles of eTM sequences in a wide range of biological processes in plants, an increasing need for discovery and annotation of previously un-annotated eTMs via computational techniques arose. However, bioinformatics strategies that take advantage of the full potential of public transcriptome data sets for discovering eTM transcripts are still lacking. Herein, I aimed to demonstrate a computational analysis pipeline for the prediction of putative eTM sequences from published public plant transcriptome libraries. Using two transcriptome libraries as an example case, I found previously un-annotated *Prunus persica* (peach) eTMs and proposed an eTM-miRNA-mRNA interaction network module for peach fruit organ development. The study predicted 44 novel eTM sequences from the libraries and a previously published set of lncRNAs, and I further found eTM related pathways by incorporating miRNA target information and annotation from previous studies. I also provided a step-by-step guideline and sample codes for predicting previously un-annotated eTM sequences from the raw transcriptome libraries of plant species.

2 Materials and Workflow

2.1 Hardware

The eTM discovery pipeline was run on Linux distribution, CentOS (RedHat) 6.x. (<https://www.centos.org/>), with 16 × 2.6 GHz (Intel E5-2650v2) CPUs and 64 GB memory.

2.2 Data Analysis

In each data analysis step, I first provide necessary command to properly run individual tools, then specify my sample code(s) starting with \$ symbol, which denotes a Linux shell prompt. Once the required tools are installed, users can simply write and execute the sample commands in Linux terminal screen to reproduce the same outputs obtained in the proposed analysis pipeline (Figure 1). All transcriptome analysis tools used in the pipeline are publicly available and they can be freely downloaded from their web sites. Users can also modify the arguments represented with < ... > symbols to customize the analysis steps.

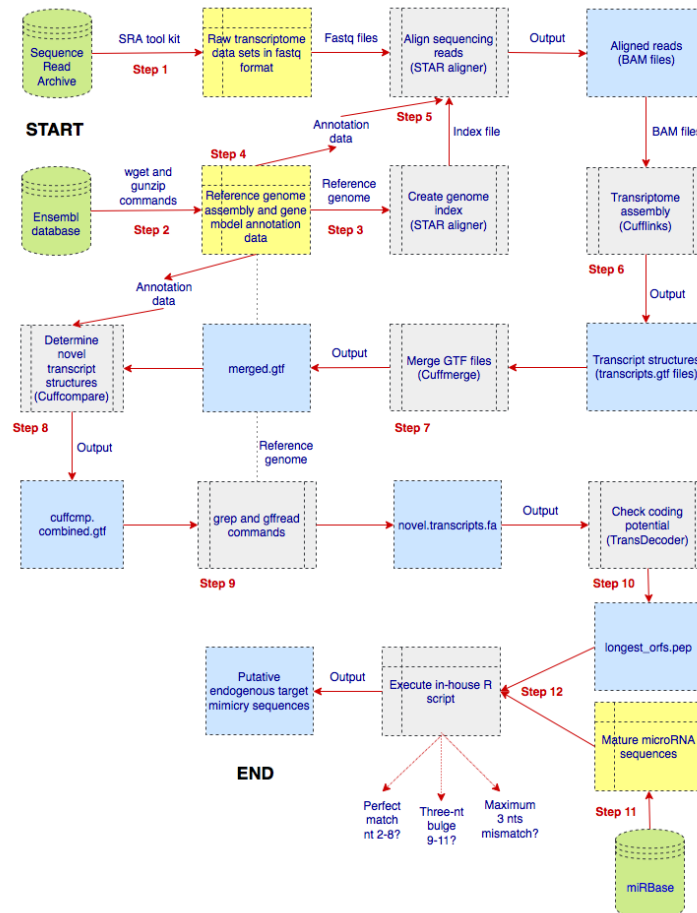


Figure 1: Schematic diagram of the study workflow.

i) In the first step, I initially need Sequence Read Archive (SRA) Tool Kit (v2.6.3) to download peach transcriptome libraries from SRA database (<http://www.ncbi.nlm.nih.gov/sra>) [30], which is a public repository of unprocessed next generation sequencing data sets. SRA Tool Kit can be downloaded from <https://github.com/ncbi/sra-tools> and its user manual can be viewed at <http://www.ncbi.nlm.nih.gov/books/NBK158900>. Herein, I utilize two transcriptome libraries, as an example case, previously generated by [31], (SRA accession IDs: SRR1300787 and SRR1300788) to identify eTM sequences, and raw sequencing files of the libraries in FASTQ format can be obtained from SRA using the following commands:

SRA Tool Kit command:

```
fastq-dump --gzip --skip-technical --readids --dumpbase --clip <SRA accession ID>
```

Sample code:

```
$ fastq-dump --gzip --skip-technical --readids --dumpbase --clip SRR1300787
```

```
$ fastq-dump --gzip --skip-technical --readids --dumpbase --clip SRR1300788
```

ii) Here, I align short reads in the FASTQ files to peach reference genome with splice-aware mapper, STAR (v2.5.1b) [32]. The source code and user manual of STAR aligner are freely available at <https://github.com/alexdobin/STAR>. I need to obtain peach DNA sequence as a zipped (.gz) multi-line fasta format (.fa) from the Ensembl ftp site. To perform download task, I first utilize the *wget* command line utility of Linux environment. Then, I make use of *gunzip* command to uncompress the downloaded file.

Sample code:

```
$ wget ftp://ftp.ensemblgenomes.org/pub/release-31/plants/fasta/prunus_persica/dna/Prunus_persica.Prupe1_0.31.dna.genome.fa.gz
```

```
$ gunzip Prunus_persica.Prupe1_0.31.dna.genome.fa.gz
```

iii) Build a STAR index (.index) that will be utilized during short read alignment to the reference genome. Herein, I use the multi-line DNA fasta file (downloaded in step ii) as an input for the STAR genomeGenerate command to generate STAR indexes.

STAR genomeGenerate command:

```
STAR --runMode genomeGenerate --genomeDir <output index folder name> --genomeFastaFiles <Assembly name .fa>
```

Sample code:

```
$ STAR --runMode genomeGenerate --genomeDir ./ppersica.star.index/ --genomeFastaFiles Prunus_persica.Prupe1_0.31.dna.genome.fa
```

iv) Download and extract the gene transfer file (.GTF), including exon structures of known transcripts annotated in the Ensembl plant database using `wget` and `gunzip` commands, respectively (as in step ii). GTF file can be downloaded using the following command:

Sample code:

```
$ wget ftp://ftp.ensemblgenomes.org/pub/release-31/plants/gtf/prunus_persica/Prunus_persica.Prupe1_0.31.gtf.gz
```

```
$ gunzip Prunus_persica.Prupe1_0.31.gtf.gz
```

v) Align short reads to the reference genome using the command below. It will generate a Binary Sequence Alignment/Map output file (.BAM), which contains read alignment data. To run STAR aligner, you will utilize the FASTQ files (downloaded in step i), STAR index (generated in step iii), and GTF file (downloaded in step v).

STAR align command:

```
STAR --genomeDir <STAR index folder name> --sjdbGTFfile <Assembly name .gtf> --readFilesIn <Sample name_1.fastq> <Sample name_2.fastq> --outFileNamePrefix <output file prefix name>
```

Sample code:

```
$ STAR --genomeDir ./ppersica.star.index/ --genomeLoad NoSharedMemory --sjdbGTFfile Prunus_persica.Prupe1_0.31.gtf --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate --twopassMode Basic --readFilesIn SRR1300787_1.fastq.gz SRR1300787_2.fastq.gz --outFileNamePrefix SRR1300787. --outFilterIntronMotifs RemoveNoncanonical
```

```
$ STAR --genomeDir ./ppersica.star.index/ --genomeLoad NoSharedMemory --sjdbGTFfile Prunus_persica.Prupe1_0.31.gtf --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate --twopassMode Basic --readFilesIn SRR1300788_1.fastq.gz SRR1300788_2.fastq.gz --outFileNamePrefix SRR1300788. --outFilterIntronMotifs RemoveNoncanonical
```

vi) The previous step generates two bam files with `Aligned.sortedByCoord.out.bam` extension in the STAR output folders. I utilize these files to perform genome-guided assembly with the popular transcriptome analysis suite, Cufflinks (v2.2.1) [10], which is available at <http://cole-trapnell-lab.github.io/cufflinks/>. Cufflinks will output transcript information in a file called “`transcripts.gtf`” for each bam file.

Cufflinks command:

```
cufflinks -g <Assembly name .gtf> <accepted_hits.bam>
```

Sample code:

```
$ cufflinks -g Prunus_persica.Prupe1_0.31.gtf --library-type fr-firststrand
```

```
SRR1300787.Aligned.sortedByCoord.out.bam
```

```
$ cufflinks -g Prunus_persica.Prupe1_0.31.gtf --library-type fr-firststrand
```

```
SRR1300788.Aligned.sortedByCoord.out.bam
```

vii) Create a text file named “`gtfs.path.txt`”, which includes the path and filename of each “`transcripts.gtf`” (generated in step vi), and then merge each “`transcripts.gtf`” files into a single one using `cuffmerge` utility of Cufflinks suite with the following command:

Cuffmerge command:

```
cuffmerge <Gtf file list.txt >
```

Sample code:

```
$ cuffmerge pperica.gtfs.path.txt
```

viii) Query the transcripts being found in the previous step against the Ensembl plant database (<http://plants.ensembl.org>) with the Cuffcompare tool, which is an analysis module of Cufflinks tool. The “`merged.gtf`” file generated with Cuffmerge (in step vii) contains exon structures of all possible transcripts, and Cuffcompare will help us to classify those transcripts as novel or known. For additional information about transcript classification, I recommend reading the Cuffcompare user manual, accessible at <http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/>. In my example, I am seeking to identify previously un-annotated intergenic transcripts; therefore I am only interested in the transcripts with class_code “`u`” within Cuffcompare output, “`cuffcmp.combined.gtf`”.

Cuffcompare command:

```
cuffcompare -r <Assembly name .gtf> <merged.gtf>
```

Sample code:

```
$ cuffcompare -r Prunus_persica.Prupe1_0.31.gtf merged.gtf
```

ix) Retrieve intergenic previously un-annotated transcript sequences with the `gffread` utility of Cufflinks. Parse “`cuffcmp.combined.gtf`”, generated in the previous step, and extract the exon structure information of the transcripts with annotated “`u`” class code with the `grep` command line in the Linux terminal window. This

command helps to process the “cuffcmp.combined.gtf” file line by line, and extracts the lines that match a specified pattern. After extracting the exon structure information of previously un-annotated transcripts from “cuffcmp.combined.gtf”, I will create a separate .gtf file named “novel.transcripts.gtf” that will be utilized for retrieving transcript sequences.

Parsing Script:

```
$ grep 'class_code "u";' cuffcmp.combined.gtf > novel.transcripts.gtf
```

gffread command:

```
gffread -w <gffread output name .fa> <Assembly name .fa> <novel.transcripts.gtf>
```

Sample code:

```
$ gffread -w novel.transcripts.fa -g Prunus_persica.Prupe1_0.31.dna.genome.fa novel.transcripts.gtf
```

x) Detect potential coding regions within the transcript sequences being extracted in the previous step.

In order to complete this step, I will utilize a computational tool called TransDecoder (v2.0.1), which can be downloaded from <https://transdecoder.github.io/>. One of the features of TransDecoder is to allow users to define the minimum length of the open reading frame (ORF), which will be searched within each transcript. In this analysis pipeline, however, I will search only the ORFs at least 50 amino acids length.

TransDecoder command:

TransDecoder.LongOrfs -m <the length of amino acids searched within transcripts> -t <fasta file name including transcript sequences>

Sample code:

```
$ TransDecoder -m 50 -t novel.transcripts.fa
```

xi) Download and extract mature miRNA sequences from miRBase (release 21) using the following commands (as in steps ii and iv):

Sample code:

```
$ wget ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz
```

```
$ gunzip mature.fa.gz
```

xii) Execute my custom eTM identification script written in R (v3.1.0) statistical computing environment [33]. The script has been developed based on the binding rules previously defined by [21]. It classifies a given sequence as eTM if: (i) the 2nd to 8th positions at the 5' end of a miRNA sequence perfectly match to the given transcript sequence, (ii) there are three nucleotides bulges between the 9th to 11th positions at the 5' end of the miRNA sequence, and (iii) maximum 3 nucleotide mismatch (excluding bulge region) exist miRNA and its target. To successfully run the script, I need “Biostrings” (<https://bioconductor.org/packages/release/bioc/html/Biostrings.html>) and “data.table” (<https://cran.r-project.org/web/packages/data.table/index.html>) R packages. Once the installation of the R packages is completed, download my custom script from <http://tools.ibg.deu.edu.tr/ppersica/eTM.search.script.zip>, and put the R script in same folder with “mature.fa” (downloaded in step xi), “novel.transcripts.fa” (generated in step ix) and “longest_orfs.pep” (generated in step x) files that have been generated in the previous steps then execute the following command:

Sample code:

```
$ Rscript eTM.search.R
```

xiii) The previous steps yield a .csv file including miRNA-eTM interaction pairs and the run time of the command above depends on the number of miRNA and transcript sequences in “mature.fa” and “novel.transcripts.fa” files.

3 Results and Discussion

One of the major advantages of RNA sequencing technique over microarray technology is to enable to discover novel transcript sequences previously remained un-annotated in public genomic databases. Utilizing diverse open-source bioinformatics tools and my custom script developed in R language, I found 11 putative eTM sequences from two paired-end peach fruit transcriptome libraries. In addition to these public data sets, I also collected all available peach lncRNA sequences from GReeNC database [34], and searched for putative eTMs with the custom R script. Consequently, I identified additional 33 putative eTMs from GreenC data set. In total, the study predicted 44 novel eTM sequences from the libraries and a previously published set of peach fruit lncRNAs. All putative eTM sequences and their predicted miRNA targets are available at <http://tools.ibg.deu.edu.tr/ppersica/eTM.sequences.and.features.xlsx>.

Computational strategies to identify previously un-annotated eTMs have a great potential to improve our understanding of the miRNA-mediated regulatory programs in plants. Based on the results of my computational analysis, I observed that one of the most widely expressed miRNA families among diverse plant species, miR156 [35], might be potentially sponged by seven putative eTMs (Figure 2). Additionally, I superimposed

the miR156 targets recently introduced by [36] with the predicted eTMs, and built an eTM-miRNA-mRNA regulatory network module (Figure 2) to demonstrate how predicted eTMs can be linked to GO terms. Per the proposed network module, two potential targets of miR156, ppa021582m and ppa007056m, were found to be involved in biological processes associated with organ development (GO:0048513) and system development (GO:0048731). Another target of miR156, ppa011968m, is predicted to be involved in multicellular organismal development (GO:0007275) [36]. This regulatory network module suggests that eTMs potentially play roles in the regulation of development processes in peach fruit via targeting specific miRNAs. Additionally, the investigation of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [37] of miRNA-targeted genes [38] revealed that the predicted eTMs might have regulatory roles in distinct metabolic processes such as amino acid and carbohydrate metabolism, and zeatin biosynthesis via targeting miRNAs involved in those pathways (Table 1).

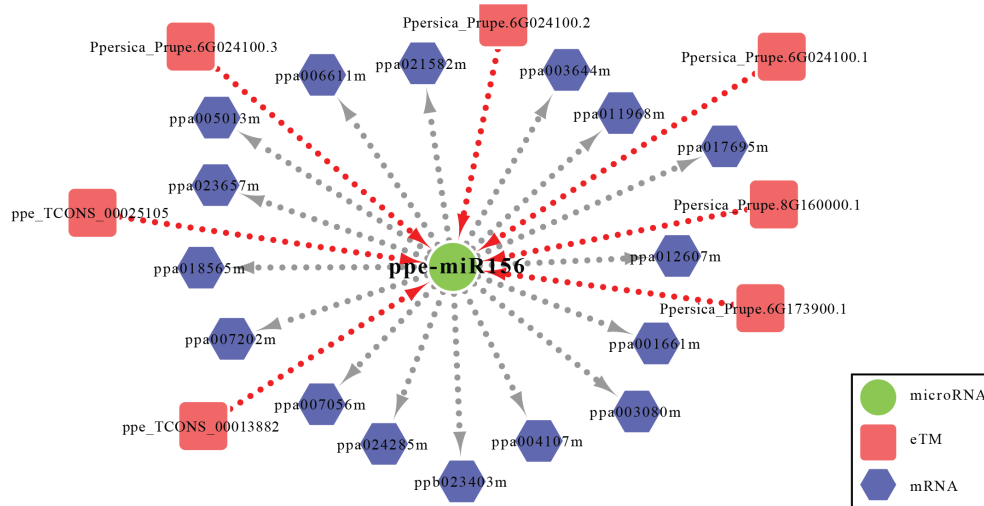


Figure 2: Computationally predicted eTM-miRNA-mRNA regulatory network module for the development of peach fruit. ETMs, mRNAs and miR156 are represented by red, blue and green colored nodes, respectively. In the proposed gene network, while the red dotted lines show putative eTM-miRNA interactions, the grey ones indicate potential relationships between miR156 and its target mRNAs.

Table 1: KEGG pathways potentially regulated by ETM-targeted miRNAs.

eTM ID	miRNA name	Associated KEGG pathways and IDs
ppe_TCONS_00013882, ppe_TCONS_00025105, Ppersica_Prupe.6G024100.1, Ppersica_Prupe.6G024100.2, Ppersica_Prupe.6G024100.3, Ppersica_Prupe.8G160000.1, Ppersica_Prupe.6G173900.1	miR156	N-acetylated-alpha-linked acidic dipeptidase [EC:3.4.17.21] (K01301)
ppe_TCONS_00032888, ppe_TCONS_00032889, Ppersica_Prupe.7G060600.1	miR159	cytokinin dehydrogenase (K00279), myb proto-oncogene protein, plant (K09422), mannan endo-1,4-beta-mannosidase (K19355)
ppe_TCONS_00029188 Ppersica_Prupe.4G246300.1, Ppersica_Prupe.4G246300.2	miR160 miR164	NA cytokinin dehydrogenase (K00279), myb proto-oncogene protein, plant (K09422), mannan endo-1,4-beta-mannosidase (K19355)
Ppersica_Prupe.4G072900.1	miR166	homeobox-leucine zipper protein (K09338)

ppe_TCONS_00000008	miR172	serine/threonine-protein phosphatase 2A regulatory subunit A (K03456), flavonoid 3'-monooxygenase (K05280), AP2-like factor, euAP2 lineage (K09284), ubiquitin carboxyl-terminal hydrolase 7 (K11838)
ppe_TCONS_00033793, Ppersica_Prupe.5G136000.1	miR394	glyceraldehyde 3-phosphate dehydrogenase (K00134), DNA-directed RNA polymerase III subunit RPC2 (K03021), 14-3-3 protein epsilon (K06630), myb proto-oncogene protein, plant (K09422), centromere protein O (K11507)
Ppersica_Prupe.5G068700.4	miR397	spermidine synthase (K00797), laccase (K05909), Ras homolog gene family, member T1 (K07870), voltage-dependent anion channel protein 2 (K15040)
Ppersica_Prupe.6G024100.1, Ppersica_Prupe.6G024100.2, Ppersica_Prupe.6G024100.3	miR482	NA
Ppersica_Prupe.1G248100.1, Ppersica_Prupe.2G063100.1, Ppersica_Prupe.4G065000.1, Ppersica_Prupe.4G160600.1, Ppersica_Prupe.4G278000.1	miR5225	coatamer, subunit beta (K17301)
Ppersica_Prupe.1G145200.1	miR530	EREBP-like factor (K09286), myb proto-oncogene protein, plant (K09422), ubiquitin carboxyl-terminal hydrolase 7 (K11838)
Ppersica_Prupe.4G267300.1	miR827	ERO1-like protein alpha (K10950)

A single miRNA can be targeted by multiple putative eTM transcripts, and involved in several metabolic pathways.

Since from the first discovery of plant eTM, *Induced by Phosphate Starvation 1 (IPS1)* transcript binding to phosphate starvation-induced miRNA, ath-miR399 in Arabidopsis, as an endogenous lncRNA [18], common interest on eTM-based miRNA regulation in plants. Subsequently, several eTM sequences of some conserved miRNAs were computationally identified in rice, soybean and other plant species [21], [39], [40]. Some of the plant eTMs were also utilized for functional characterization of miRNAs [18], [26]. Overexpression of miR-156 eTM (MIM156) and miR-319 eTM (MIM319) resulted altered phenotypes in *A. thaliana* [18]. Therefore, identification of eTM sequences as first step of functional characterization of eTM-miRNA-target transcript modules will greatly help to researchers in the field. By utilization of presented workflow, it will be beneficial to discover previously un-annotated plant eTMs. On the other hand, as one of the important conserved miRNAs, miR156, with its regulatory roles on *SQUAMOSA-PROMOTER BINDING PROTEIN-LIKE (SPL)* family of transcription factors [41] for a number of vital functions such as vegetative and reproductive phase changes [42], leaf morphology and development [43], panicle architecture [44], stress responses [45], [46]. Multiple eTM sequences for miRNA156 in peach found in this study suggest that miRNA with a number of SPL family transcript targets might be under control of several specific miR156 eTMs. The identified miR156 eTMs will then be utilized for functional characterization and enhancement of agronomical traits.

In conclusion, with the help of the presented pipeline in this manuscript, I believe that previously un-annotated plant eTMs can be predicted effectively from next generation sequencing transcriptome datasets, and these eTM sequences can be utilized for the establishment of the eTM-miRNA-mRNA regulatory networks. However, researchers should note that one of the purposes of this manuscript is to provide a general overview regarding novel eTM sequence discovery, and I strongly recommend researchers to examine the user manual of each tool being used to select appropriate parameters in line with their research goals.

Acknowledgements

I would like to thank ÜnverLab for its support in implementing the analysis pipeline.

Conflict of interest statement: The author states no conflict of interest. The author has read the journal's publication ethics and publication malpractice statement available at the journal's website and hereby confirms that he complies with all its parts applicable to the present scientific work.

References

- [1] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010;28:511–5.
- [2] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol.* 2010;28:503–10.
- [3] Chaitankar V, Karakulah G, Ratnapriya R, Giuste FO, Brooks MJ, Swaroop A. Next generation sequencing technology and genomewide data analysis: perspectives for retinal research. *Prog Retin Eye Res.* 2016;55:1–31.
- [4] Bolger ME, Arsova B, Usadel B. Plant genome and transcriptome annotations: from misconceptions to simple solutions. *Brief Bioinform.* 2017;bbw135. DOI:10.1093/bib/bbw135.
- [5] Liu S, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol.* 2016;17:67.
- [6] Alvarez-Dominguez JR, Bai Z, Xu D, Yuan B, Lo KA, Yoon M, et al. De Novo reconstruction of adipose tissue transcriptomes reveals long non-coding RNA regulators of brown adipocyte development. *Cell Metab.* 2015;21:764–76.
- [7] Parikhshak NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, Gandal MJ, et al. Genome-wide changes in lincRNA, splicing, and regional gene expression patterns in autism. *Nature.* 2016;540:423–7.
- [8] Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, et al. Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* 2008;9:R175.
- [9] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 2015;33:290–5.
- [10] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 2012;7:562–78.
- [11] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics.* 2012;28:1086–92.
- [12] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29:644–52.
- [13] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 2013;8:1494–512.
- [14] Eldem V, Okay S, Ünver T. Plant microRNAs: new players in functional genomics. *Turk J Agric For.* 2013;37:1–21.
- [15] Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 2009;19:92–105.
- [16] Griffiths-Jones S. miRBase: the microRNA sequence database. *Methods Mol Biol.* 2006;342:129–38.
- [17] Gupta PK. MicroRNAs and target mimics for crop improvement. *Curr Sci India.* 2015;108:1624–33.
- [18] Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, et al. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet.* 2007;39:1033–7.
- [19] Zhang H, Chen X, Wang C, Xu Z, Wang Y, Liu X, et al. Long non-coding genes implicated in response to stripe rust pathogen stress in wheat (*Triticum aestivum* L.). *Mol Biol Rep.* 2013;40:6245–53.
- [20] Meng M, Shao C, Wang H, Jin Y. Target mimics: an embedded layer of microRNA-involved gene regulatory networks in plants. *BMC Genomics.* 2012;13:197.
- [21] Wu HJ, Wang ZM, Wang M, Wang XJ. Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants. *Plant Physiol.* 2013;161:1875–84.
- [22] Gupta PK. MicroRNAs and target mimics for crop. *Curr Sci.* 2015;108:1624–33.
- [23] Todesco M, Rubio-Somoza I, Paz-Ares J, Weigel D. A collection of target mimics for comprehensive analysis of microRNA function in *Arabidopsis thaliana*. *PLoS Genet.* 2010;6:e1001031.
- [24] Karakulah G, Yucebilgili Kurtoglu K, Ünver T. PeTmBase: a database of plant endogenous target mimics (eTMs). *PLoS One.* 2016;11:e0167698.
- [25] Yuan C, Meng X, Li X, Illing N, Ingle RA, Wang J, et al. PceRBase: a database of plant competing endogenous RNA. *Nucleic Acids Res.* 2017;45:D1009–D14.
- [26] Li F, Wang W, Zhao N, Xiao B, Cao P, Wu X, et al. Regulation of nicotine biosynthesis by an endogenous target mimicry of microRNA in tobacco. *Plant Physiol.* 2015;169:1062–71.
- [27] Ye CY, Xu H, Shen E, Liu Y, Wang Y, Shen Y, et al. Genome-wide identification of non-coding RNAs interacted with microRNAs in soybean. *Front Plant Sci.* 2014;5:743.
- [28] Bonnet E, He Y, Billiau K, Van de Peer Y. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics.* 2010;26:1566–8.
- [29] Wang P, Zhi H, Zhang Y, Liu Y, Zhang J, Gao Y, et al. miRSponge: a manually curated database for experimentally supported miRNA sponges and ceRNAs. *Database (Oxford).* 2015;bav0982015.
- [30] Leinonen R, Sugawara H, Shumway M. International Nucleotide Sequence Database C. The sequence read archive. *Nucleic Acids Res.* 2011;39:D19–21.

- [31] Bakir Y, Eldem V, Zararsiz G, Unver T. Global transcriptome analysis reveals differences in gene expression patterns between nonhyperhydric and hyperhydric peach leaves. *Plant Genome*. 2016;9.
- [32] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
- [33] Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0 2014.
- [34] Paytuvi Gallart A, Hermoso Pulido A, Anzar Martinez de Lagran I, Sanseverino W, Aiese Cigliano R. GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Res*. 2016;44:D1161–6.
- [35] Axtell M, Bartel DP. Antiquity of microRNAs and their targets in land plants. *Plant Cell*. 2005;17:1658–73.
- [36] Zhang C, Zhang B, Ma R, Yu M, Guo S, Guo L, et al. Identification of known and novel microRNAs and their targets in peach (*Prunus persica*) fruit by high-throughput sequencing. *PLoS One*. 2016;11:e0159253.
- [37] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 1999;27:29–34.
- [38] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. 2007;35:W182–5.
- [39] Ye C, Xu H, Shen E, Liu Y, Wang Y, Shen Y, et al. Genome-wide identification of non-coding RNAs interacted with microRNAs in soybean. *Front Plant Sci*. 2014;5:743.
- [40] Banks IR, Zhang Y, Wiggins BE, Heck GR, Ivashuta S. RNA decoys: an emerging component of plant regulatory networks?. *Plant Signal Behav*. 2012;7:1188–93.
- [41] Wang H, Wang H. The miR156/SPL module, a regulatory hub and versatile toolbox, gears up crops for enhanced agronomic traits. *Mol Plant*. 2015;8:677–88.
- [42] Poethig RS. Vegetative phase change and shoot maturation in plants. *Curr Top Dev Biol*. 2013;105:125–52.
- [43] Wu G, Poethig RS. Temporal regulation of shoot development in *Arabidopsis thaliana* by miR156 and its target SPL3. *Development*. 2006;133:3539–47.
- [44] Jiao Y, Wang Y, Xue D, Wang J, Yan M, Liu G, et al. Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nat Genet*. 2010;42:541–4.
- [45] Eldem V, Akçay UÇ, Ozhuner E, Bakır Y, Uranbey S, Unver T. Genome-wide identification of miRNAs responsive to drought in peach (*Prunus persica*) by high-throughput deep sequencing. *PLoS One*. 2012;7:e50298.
- [46] Cui N, Sun X, Sun M, Jia B, Duanmu H, Lv D, et al. Overexpression of OsmiR156k leads to reduced tolerance to cold stress in rice (*Oryza Sativa*). *Mol Breed*. 2015;35:1–11.