psRobot manual

DESCRIPTION

psRobot is a smRNA analysis software package, which so far contains four functions: (1) psRobot_map is designed to find all the perfect matching locations of short sequences (less than 40bp) in longer reference sequences; (2) psRobot_mir is designed to find potential miRNAs or small hairpin RNAs (shRNAs) from high throughput sequencing data. (3) psRobot_tar is designed to find potential small RNA targets on a large scale. (4) psRobot_deg is designed to identify which small RNA targets are supported by user specified degradome data.

RELEASE

Release	Data	Changes
Version 1.1	05/29/2012	The first public release
Version 1.2	06/21/2012	use a relatively strict criterion to detect miRNA* sequneces (see manual of psRobot_mir below). Fix a bug in psRobot_mir that could cause some miRNA*s to have identical IDs. update scripts to be more compatible in different Linux distributions.

CITATION

Wu HJ, Ma YK, Chen T, Wang M, Wang XJ. (2012) PsRobot: a web-based plant small RNA meta-analysis toolbox. Nucleic Acids Res. DOI:10.1093/nar/gks554. (Full Text)

REQUIREMENTS

perl 5 or greater

gcc 4.0 or greater

psRobot_mir requires “nafold” command in **Mfold-3.5** software to get the potential folding structure of precursor RNAs.

INSTALLATION

The simplest way to install this package is (require root permission):

./configure

make

make install (log as root)

Install psRobot in an alternative path with no root permission:

./configure -p <your full path> -l <your full path>

make

make install

Configure options:

-h	display this help and exit
-p PREFIX	install architecture-independent files in PREFIX [complete path needed, default /usr/bin]
-l LIBS	install perl libs in LIBS [complete path needed, default standard perl lib, the path next to last in your @INC]

COMMANDS AND OPTIONS

psRobot_map **psRobot_map <*short_sequences*> <*reference_sequences*> <*output*>**

<*short_sequences*> is the first input file which contains the short sequences seperated by <Tab>:

<ID>	<SEQ> **(Don't include this row!)**
SrID001	TGACAGAAGAGAGTGAGCAC
SrID002	TCGCTTGGTGCAGATCGGGAC
SrID003	TTGACAGAAGAGAGTGAGCAC
SrID004	TGAAGCTGCCAGCATGATCTGA
SrID005	GGCGGATGTAGCCAAGTGGA

(Don't use "blank" in your short sequence IDs!)

<*reference_sequences*> is the second input file which contains the reference sequences in fasta format. **(Don't use blank in your reference sequence IDs!)**

<*output*> is the output file containing the mapping results:

ID	ref	strand	start	end	seq
SrID003	ref01	+	28	48	TTGACAGAAGAGAGTGAGCAC
SrID001	ref01	+	29	48	TGACAGAAGAGAGTGAGCAC
SrID001	ref01	+	60	79	TGACAGAAGAGAGTGAGCAC

Field	Description
ID	Short sequence ID in <*short_sequences*>
ref	Reference sequence ID in <*reference_sequences*>
strand	Mapping strand
start	Start position
end	End position
seq	Short sequence

psRobot_mir

psRobot_mir -s <*smRNA*> -g <*genome*>

Note: Different psRobot_mir run must be in **different folders**.

Options:

-s input file name which contains smRNA deep sequencing data in tab-delimited (tsv) format. [default: smRNA]

Suggested format:

<smRNA sequence><Tab><reads1><Tab><reads2><Tab><reads3>…

The first column must be smRNA sequences. The following columns are the smRNA clone counts in various samples/conditions. Columns must be seperated by <Tab>.

-g input file name which contains reference genome/contig sequences in FASTA format. [default: genome]

-k input file name which contains known miRNA GFF3 file corresponding to the reference genome (GFF3 format). The reference genome IDs must be identical to the IDs used in reference genome sequence file (-g file). This file is needed only If one wants to exclude known miRNAs from prediction results. [default: kmiRNA]

-r clone counts selection: minimal smRNA clone counts. SmRNAs with less than this clone counts will be excluded from the following analysis. [default: 2]

-l loci selection: the maximal number of smRNA mapping locations in reference genome. SmRNAs with more than this number of mapping locations will be excluded from the following analysis. [default: 20]

-cg minimal number of nucleartides between adjacent smRNA clusters. [default: 200]

-cl maximal length limitation of smRNA clusters selected to predict stem-loop precursors. [default: 300]

-cr all the clone counts of smRNAs in one cluster will be summed up to represent the smRNA cluster's clone counts. SmRNA clusters with more than this clone counts will be selected to predict stem-loop precursors. [default: 10]

-mr the highest expressed smRNA in one cluster higher than this clone counts will be selected to predict stem-loop precursors. [default: 10]

-mml minimal number of mismatches in supposed miRNA mature region. [default: 1]

-mmh maximal number of mismatches in supposed miRNA mature region. [default: 5]

-ll retain large loop miRNAs (T/F). [default: F]

Output:

<***.StarInfo***> contians the major information of predicted miRNAs:

Col_1	Col_2	Col_3	Col_4	Col_5	Col_6	Col_7
Sr7_1_1_21_188_60	CTGAAGTGTTTGGGGGGACTC	2	Sr4_1_1_21_3_3	gttccctttaacgcttcattg	ref01:20:132:+	attgacttcaaaaaatatgagttccctttaacgcttcattgttgaatactcaa
Sr9_1_1_21_2000_10	AGAAAACTTTCTGGAGACCAA	2	Sr5_1_1_21_30_1	tggtcttcagaaagttttcttg	ref02:708:910:-	gaggagcaagccttgatgattcgacaaagtgaagggtttggtcttcagaaagt
Sr3_1_1_21_200_100	ACATATGATCTGCATCTTTGCATT	2	Sr1_1_1_7_3	tgcaaagatgcagatcatatgtcc	ref03:6953:7055:+	atttcttgaggcttttgataaacatggACATATGATCTGCATCTTTGCATTttc
Sr12_1_1_21_9_33	AAAGATGCAGATCATATGTCC	2	Sr13_1_1_21_2_7	acatatgatctgcatctttgc	ref04:6972:7034:-	atggacatatgatctgcatctttgcacttgaaaatgcAAAGATGCAGATCATA
Sr2_1_1_21_9_14	AACTAATTTTTATTTGGACGTTTA	2	-	-	ref04:11:173:-	atatacgggaaaaaactcgaaaaaacctcAACTAATTTTTATTTGGACGTTTA

Col	Description
1	Information of mature miRNA. Format: [miRNA ID] _ [No. of miRNA mapping loci] _ [order of miRNA mapping location] _ [miRNA length] _ [Reads1] _ [Reads2]…
2	miRNA sequence
3	No. of smRNA clusters in miRNA precursor. The canonical miRNAs always have two clusters on their precursors, one is at miRNA mature region and the other is at miRNA* region.
4$	Information of miRNA*. "-" represents no miRNA* sequence can be detected. Format: [miRNA* ID] _ [No. of miRNA* mapping loci] _ [order of miRNA* mapping location] _ [miRNA* length] _ [Reads1] _ [Reads2]…
5$	miRNA* sequence. "-" represents no miRNA* sequence can be detected.
6	Location of miRNA precursor in reference genome. Format: [reference genome ID] _ [start] _ [end] _ [strand]
7	miRNA precursor sequence

$: miRNA* is typically derived from the opposite stem-arm of the miRNA precurosr with 2-nt 3' overhang compared to mature miRNA. However, the 2-nt 3' overhang is sometimes too ideal for miRNA* detection. Therefore, ±3nt shift from the expected ideal miRNA* position is permitted in psRobot_mir's miRNA* detection.

<***.Struc***> contians the stem-loop folding structures of predicted miRNA precursors.

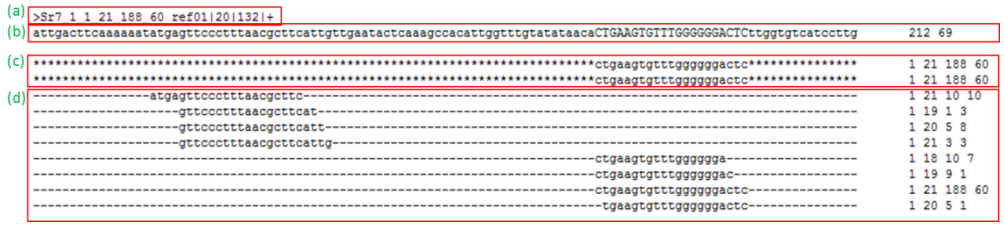
```
Sequence      1 Structure     1

Folding bases       1 to     113 of Sr7_1_1_21_188||77||15 (a)
 dG =    -44.30 (b)

           10        20        30        40        50        (c)
at----|    t    aaaatat         tt         t     a    tca     ac
      tgac tcaa       gagttccct  aacgcttca tgttg atac   aagcc  \
      actg ggtt       CTCAGGGGG  TTGTGAAGT acaat tatg   tttgg  a
gttcct^    t    -------          GT         C     a    ---     tt
 110        100               90        80        70
```

Field	Description
(a)	Information of mature miRNA. Format: [miRNA ID] _ [No. of miRNA mapping loci] _ [order of miRNA mapping location] _ [miRNA length] _ [Max Reads in all samples] ‖ [No. of upstream nucleotides to consist precursor] ‖ [No. of downstream nucleotides to consist precursor]
(b)	Free energy of the precursor folding structure
(c)	Precursor folding structure

<***.Reads***> contains the small RNA reads distribution on predicted miRNA precursors.



Field	Description
(a)	Information of mature miRNA and genomic location of miRNA precursor. Format: >[mature miRNA]<*Tab*>[precursor location] Format of [mature miRNA]: [miRNA ID] _ [No. of miRNA mapping loci] _ [order of miRNA mapping location] _ [miRNA length] _ [Reads1] _ [Reads2]… Format of [precursor location]: [chromosome_ID] _ [start] _ [end] _ [strand]
(b)	Precursor sequence and total clone counts of mature miRNA and its ±3-nt variants
(c)	The highest expressed small RNAs on each precursor in different libraries. The numbers in the right represents [No. of genomic locations] [length of smRNA] [Reads1] [Reads2]…
(d)	Small RNAs distribution of each precursor sequence. The numbers in the right represents [No. of genomic locations] [length of smRNA] [Reads1] [Reads2]…

psRobot_tar

psRobot_tar -s <*smRNA*> -t <*target*> -o <*smRNA-target.gTP*>

Scoring rules:

1. Mismatches, gaps or bulges are evaluated with a penalty of **+1**.
2. G:U pairs are evaluated a penalty of **+0.5**.
3. **Full** penalty scores will be evaluated in user defined **essential sequence region**, and **half** penalty scores will be evaluated **outside** of essential sequence region.

Options:

-s input file name which contains smRNA sequences in FASTA format. [default: smRNA]

-t input file name which contians target sequences in FASTA format as prediction library. [default: target]

-o output file name. [default: smRNA-target.gTP]

-ts target penalty score threshold (0-5, **lower is better**). [default: 2.5]

-fp 5 prime boundary of essential sequence (1-2). [default: 2]

-tp 3 prime boundary of essential sequence (15-31). [default: 17]

-gl position after which with gap/bulge permit (0-30, 0 means no gap permitted). [default: 17]

-p number of processors to use. [default: 1]

-gn number of gaps/bulges permitted (0-5). [default: 1]

Output:

<smRNA-target.gTP> is the output file containing the target prediction results:

```
>smRNA01        Score: 1.0      tar02

Query:          1 TGACAGAAGAGAGTGAGCAC 20
                  ||||||||||||*||||||
Sbjct:       2387 ACTGTCTTCTCTCTCTCGTG 2368


>smRNA02        Score: 0.8      tar01

Query:          1 TGCCAAAGGAGATTTGCCCTG 21
                  ||||||||||||||||*||:
Sbjct:        963 ACGGTTTCCTCTAAACGAGAT 943
```

psRobot_deg **psRobot_deg <*degradome_data*> <*target_sequences*> <*smRNA-target.gTP*><*output*>**

Note: psRobot_deg needs the **direct output of psRobot_tar**. So one should run psRobot_tar to get the .gTP file or generate your own .gTP file first.

<degradome_data> is the first input file which contains the degradome sequences data in tab-delimited (tsv) format:

<Sequence>	<Counts> **(Don't include this row!)**
TGACAGAAGAGAGTGAGCAC	108
TCGCTTGGTGCAGATCGGGAC	373
TTGACAGAAGAGAGTGAGCAC	10

<target_sequences> is the second input file which contains the target sequences in FASTA format as prediction library and **must be identical to <*target*>** file in psRobot_tar.

<smRNA-target.gTP> is the third input file containing the target prediction results, which could be the direct output of psRobot_tar.

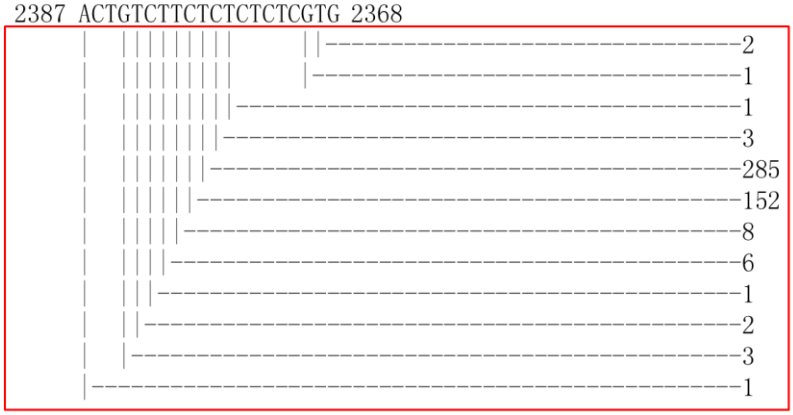
<output> is the output file containing the degradome data supported smRNA target prediction results:

(a)

```
>ath_miR156a     Score: 1.0      Deg: 285:2378:285:1635  AT1G27370.1

Query:           1 TGACAGAAGAGAGTGAGCAC 20
                   ||||||||||||*||||||
Sbjct:       2387 ACTGTCTTCTCTCTCTCGTG 2368
```

(b)



Field	Description
(a)	Degradome support. Format: Deg: [reads of predicted cleavage site] **:** [position of predicted cleavage site on target sequence] **:** [maximal reads of all cleavage sites on target sequence] **:** [total reads of all cleavage sites on target sequence]
(b)	Degradome sequence distribution on smRNA target sites. Number on each row represents the reads of degradome sequence on each cleavage site (**cleaved on the right-hand of the pointing nucleotide**).