

Manual for tsRFinder

Qinhu Wang and Weixing Shan*

Northwest A&F University

July 16, 2014

Contents

1	Introduction	1
2	How to install	1
2.1	Dependencies	1
2.2	Installation	2
3	How to use	3
3.1	Preparation the dataset	3
3.2	Running the pipeline	3
4	Demo	4
4.1	Demo data	4
4.2	Demo running	4
4.3	Demo output	7
4.4	Visualization of tmap data	8
5	FAQ	10

*Email: wxshan@nwafu.edu.cn

1 Introduction

Small RNAs are key regulators of gene expression, such as miRNA, siRNA, and piRNA. The tRNA-derived small RNA (tsRNA; also called tRNA-derived RNA fragments, tRF), is a class of novel small RNA have been identified recently. However, there is no public tool available for tsRNA analysis yet. We thus developed tsRFinder for tsRNA prediction. It takes the small RNA sequencing data and the reference genome sequence, and identifies tsRNA with additional sequence and statistical analysis for you automatically.

2 How to install

2.1 Dependencies

tsRFinder depends on a few programmes, please check and install them ¹ at first:

- Perl, greater than v5.10.1, required, for tsRFinder.pl execution. It is always installed already in most of the UNIX-like operating systems. Two perl modules, Config::Simple and Getopt::Std (always build-in) are need to install if you haven't.
- bowtie, greater than v1.0.0, required, for small RNA mapping.
<https://github.com/BenLangmead/bowtie>
- R, greater than v2.15.2, required, for small RNA data analysis and illustration.
<http://www.r-project.org>
- tRNAscan-SE, greater than v1.3.1, optional if you want to prepare tRNA input yourself, for tRNA prediction.
<http://lowelab.ucsc.edu/tRNAscan-SE>
- fastx_toolkit, greater than v0.0.14, optional if you have already processed the raw sequencing data.
https://github.com/agordon/fastx_toolkit

¹The dependencies version were based on the oldest test environment we have tested.

2.2 Installation

tsRFinder is maintained on GitHub and is ready-to-use, no compilation is required. However, if you take some time to improve the configuration, it may save you a lot of time for trouble shooting.

First, you can clone² tsRFinder by typing:

```
git clone https://github.com/wangqinhu/tsRFinder.git
```

in the terminal. Alternatively, you can download it from

```
https://github.com/wangqinhu/tsRFinder/releases/latest
```

and then decompress the zip file or the tarball.

When tsRFinder is cloned or unpacked, move the entire directory to an proper place (or current working directory, such as home directory) and add the tsRFinder path to your environment setting. For example, if tsRFinder is placed in /your/path/of/tsRFinder, then type the following in the terminal if you are using bash.

```
echo export tsR_dir="/your/path/of/tsRFinder" >> ~/.bashrc  
source ~/.bashrc
```

And now, you can run tsRFinder for your dataset.

3 How to use

3.1 Preparation the dataset

Before running tsRFinder, you are asked to prepare/download the following two files: (1) the reference genome sequence, or the reference tRNA sequence and, (2) the small RNA reads.

We strongly recommend you using the reference genome sequence and the raw small RNA sequencing data, since tsRFinder can help you prepare the reference tRNA data and clean small RNA data automatically. If you want to prepare the tRNA file and small RNA reads file by yourself, you can run the demo data (enable debug mode in command line option: *-m debug*) and

²If git is not installed, download it from <http://git-scm.com>

find out which exact format of tRNA reference and small RNA reads file can be accepted instead, this is allowed but not encouraged. You can also found the format description in the FAQ part of this document.

3.2 Running the pipeline

tsRFinder supplies two ways for arguments input, you can use both configuration file and command line option. We recommend you use a command line option for debugging and building your configuration file. Once your inputs have been determined, you can write it to a configuration file for your analysis.

If tsRFinder is properly installed, you can run tsRFinder from your terminal directly, see the usage below.

tsRFinder usage:

```
tsRFinder.pl <option>

-c Configuration file
-l Label
-g Reference genomic sequence
-t Reference tRNA sequence
-s Small RNA sequence
-a Adaptor sequence
-n Min read length           [default 18]
-x Max read length           [default 45]
-f Small RNA family threshold [default 72]
-w tRNA with/without label   [default no]
-m mode, run/debug           [default run]
-h Help
-v Version
```

Example:

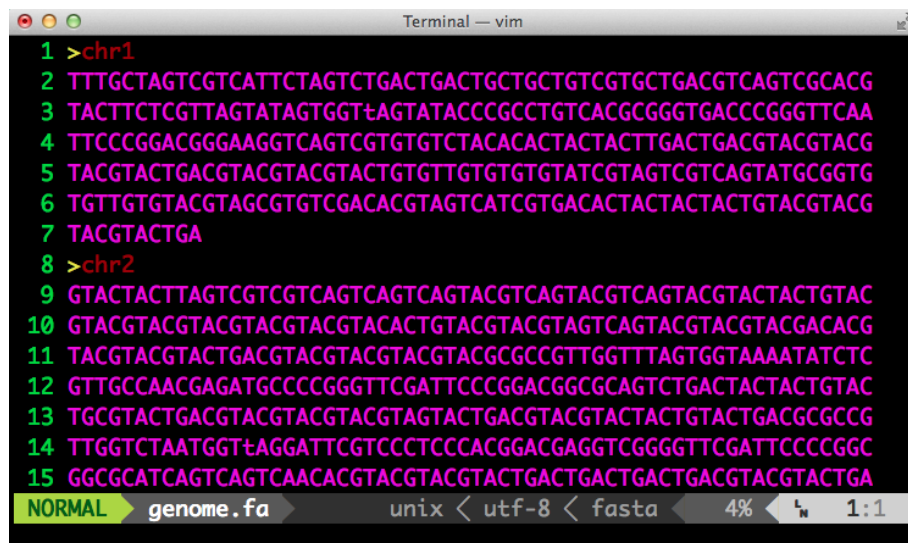
```
tsRFinder.pl -c demo/tsR.conf
```

4 Demo

To have a quick but rough overview that how tsRFinder looks like, see the the animated gif demo in doc/demo.gif or online.

4.1 Demo data

Demo refseq: we used several random sequences embedded with some real tRNAs as pseudo reference genome sequence. This small sequence data in FASTA format can be accessed in the file "tsRFinder/demo/genome.fa". In your analysis, if you have a reference genome, just replace it with the reference sequence; if you don't have reference genome sequence, you can use the reference tRNA sequence instead. Figure 1 shows what a refseq file looks like.



```
1 >chr1
2 TTTGCTAGTCGTCATTCTAGTCTGACTGACTGCTGCTGCTGCTGACGTCAGTCGCACG
3 TACTTCTCGTTAGTATAGTGGT+AGTATACCCGCCCTGTCACGCGGGTGACCCGGGTTCAA
4 TTCCCGGACGGGAAGGTCAGTCGTGTGTCTACACACTACTACTTGACTGACGTACGTACG
5 TACGTAAGTACGTACGTACGTACTGTGTTGTGTGTATCGTAGTCGTCAGTATGCGGTG
6 TGTGTGTGTACGTAGCGTGTGACACGTAGTCATCGTGACACTACTACTACTGTACGTACG
7 TACGTACTGA
8 >chr2
9 GTACTACTTAGTCGTCGTCAGTCAGTCAGTACGTACGTACGTACGTACGTACTACTGTAC
10 GTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
11 TACGTACGTACTGACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGT
12 GTTGCCAACGAGATGCCCCGGGTTGATTCCCGGACGGCGCAGTCTGACTACTACTGTAC
13 TCGTACTGACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTAC
14 TTGGTCTAATGGT+AGGATTCGTCCCTCCACGGACGAGGTCGGGGTTCGATTCCCGGC
15 GCGCATCAGTCAGTCAACACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
```

Figure 1: Screenshot of the reference sequence in fasta format

Demo sRNA: we extracted a bit of raw reads from some real experimental data as a demo here. Each read of the raw small RNA data have 4 lines, just like what we have show in Figure 2.

```

1 @HWI-ST1106:652:H0Y8BADXX:2:1101:2120:2141 1:N:0:ATGTCA
2 GAGGAGGATGACCTTGGGCTGAGAATGGAATTCTCGGGTGCCAAGGAACT
3 +
4 CCCCCFFHHHHHJJDDGGHIBGIJJICHIIJGHIIJGIJJIIJEGIIIG
5 @HWI-ST1106:652:H0Y8BADXX:2:1101:2016:2163 1:N:0:ATGTCA
6 TGAGCTACGTGGAAGACATCGTCAGGTGGAATTCTCGGGTGCCAAGGAACT
7 +
8 CCCDFADFHHHHHJGCEEHHE@GHIJ?FHIJJIIJJJJ2@AHIIGGIGB
9 @HWI-ST1106:652:H0Y8BADXX:2:1101:2346:2189 1:N:0:ATGTCA
10 TTAATATTCCTGAACCGAGACGTGGAATTCTCGGGTGCCAAGGAACTCCA
11 +
12 ???D;DD:D4D?DEIE1)3)@FFE@1@DACB9?;DDDD)?B#####
13 @HWI-ST1106:652:H0Y8BADXX:2:1101:2392:2203 1:N:0:ATGTCA
14 TCGCGCGATTGAGCGATTGAGCCTGGAATTCTCGGGTGCCAAGGAACT
15 +
16 @@@DDDDD:??FHGGHIIIGDHEHEDD9C=CGGGGHIGG<EHHHBB@DCC
17 @HWI-ST1106:652:H0Y8BADXX:2:1101:2588:2161 1:N:0:ATGTCA
18 AACGAGATGCCCCGGGTTTCGATTCGCGGACGGCGCACCATGGAATTCTCG
19 +
NORMAL sRNA.fq unix < utf-8 < fastq 0% 1:1

```

Figure 2: Screenshot of the small RNA sequence in fastq format

4.2 Demo running

tsRFinder allow you specify your inputs via a separate configuration file, for example, here is the content of our demo tsR.conf:

```

#=====#
#
# Configuration for tsRFinder
#
#=====#

mode                : run
label               : Abc
reference_genome     : demo/genome.fa
reference_tRNA       :
sRNA                : demo/sRNA.fq
adaptor             : TGGAATTCTCGGGTGCCAAGG
min_read_length     : 18

```

```
max_read_length      : 45
family_threshold     : 72
tRNA_with_label      : no
```

Currently we have 10 items (13 options in command line) used for configuration file filling. The argument items and the inputs are separated by colon (":"). We recommend you using the first three letters of the organism you are analysing as a label (e.g. for *Arabidopsis thaliana* we use *Ath*, *-l* in command line option); the paths of reference genome (*-g*) and small RNA (*-s*) should be supplied at least. If you are using a raw sequence data, please also input the adaptor sequence (*-a*). If you are not using a reference tRNA (*-t*) prepared by yourself, leave this argument to EMPTY please.

Once your configuration file is well prepared, typing the following in the terminal to run tsRFinder. In this demo, the configuration file (*-c*) is at demo/tsR.conf, so we write like this:

```
# We have set "tsR_dir" as an environment variable before
cd tsR_dir
./tsRFinder.pl -c demo/tsR.conf
```

tsRFinder can work in run and debug modes (*-m*), if you find some problem or want to extract some temporary files, you can enable debug mode, otherwise please use run mode (default mode). To check the usage or version, you can use *-h* and *-v* option, respectively.

4.3 Demo output

By default, tsRFinder will give you a summary of which files have been outputted and some basic statistics. The predicted or user inputted tRNA sequence, the small RNA clean data, the tRNA reads, and the predicted tsRNA sequence were listed. Meanwhile, tsRFinder gives you additional summary on tRNA/tsRNA expression (including 5' tsRNA and 3' tsRNA), text map (tmap) of small RNA mapped to tRNA, tsRNA family, graphics showing the expression evaluated by small RNA data, the cleavage site, and also the cleavage profile (Figure 3). Additional, tsRFinder supplies a figure showing the small RNA and tRNA reads length distribution (Figure 4) for you.

See our demo summary here:

SUMMARY

```
tRNA seq : /Users/wangqinhu/tsRFinder/Abc/tRNA.fa
  Total : 5

sRNA reads : /Users/wangqinhu/tsRFinder/Abc/sRNA.fa
  Total : 12432
  Unique : 7412

tRNA reads : /Users/wangqinhu/tsRFinder/Abc/tRNA.read.fa
  Total : 286
  Unique : 52

tsRNA seq : /Users/wangqinhu/tsRFinder/Abc/tsRNA.seq
  Total : 7
  Unique : 6

tsRNA report : /Users/wangqinhu/tsRFinder/Abc/tsRNA.report.xls

text map : /Users/wangqinhu/tsRFinder/Abc/tsRNA.tmap

visual map : /Users/wangqinhu/tsRFinder/Abc/images

distribution : /Users/wangqinhu/tsRFinder/Abc/distribution.pdf

cleavage :
  detail : /Users/wangqinhu/tsRFinder/Abc/cleavage.txt
  profile : /Users/wangqinhu/tsRFinder/Abc/cleavage_profile.pdf

tsRNA family : /Users/wangqinhu/tsRFinder/Abc/tsRNA.fam

stat. by BDI :
  Sensitivity : 0.9593
  Specificity : 0.7881
  Accuracy : 0.8847
```

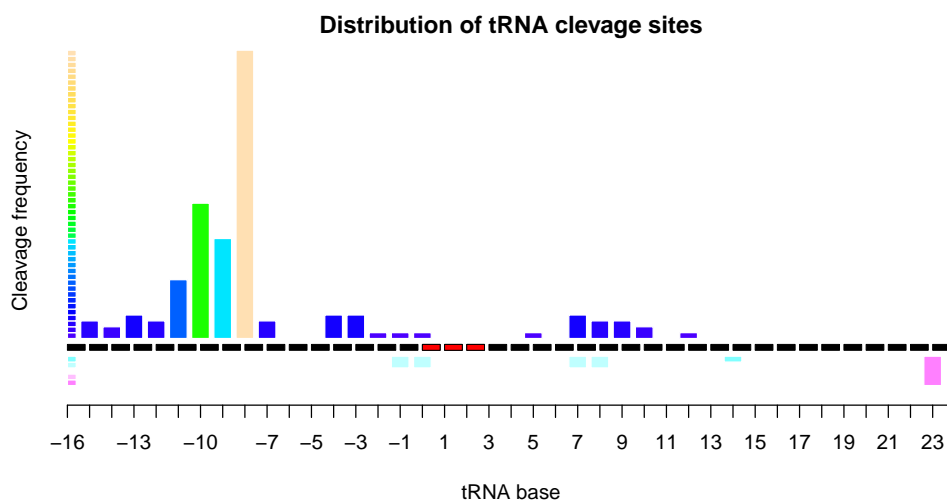


Figure 3: tRNA cleavage profile ³

4.4 Visualization of tmap data

To examine the map of tsRNA, we developed a vim syntax plugin for visualization (Figure 5). If you want to enable color text map, copy `lib/tmap.vim` into your vim syntax folder, and put the following line into your `.vimrc` file:

```
au BufNewFile,BufRead *.tmap setf tmap
```

When tmap.vim is correctly installed, open the tsRNA.tmap file with vim you will obtain a color text map, like this:

```
vim tsRNA.tmap
```

If you prefer plain text view without highlighting, just open tsRNA.tmap with any kind of text editors you have.

5 FAQ

1. Can tsRFinder run on Windows?

No. tsRFinder used some build-in program of UNIX-like systems, for example, awk, grep and head, thus running on Windows may lead unexpected errors, we strongly recommend you running tsRFinder on Linux or OS X.

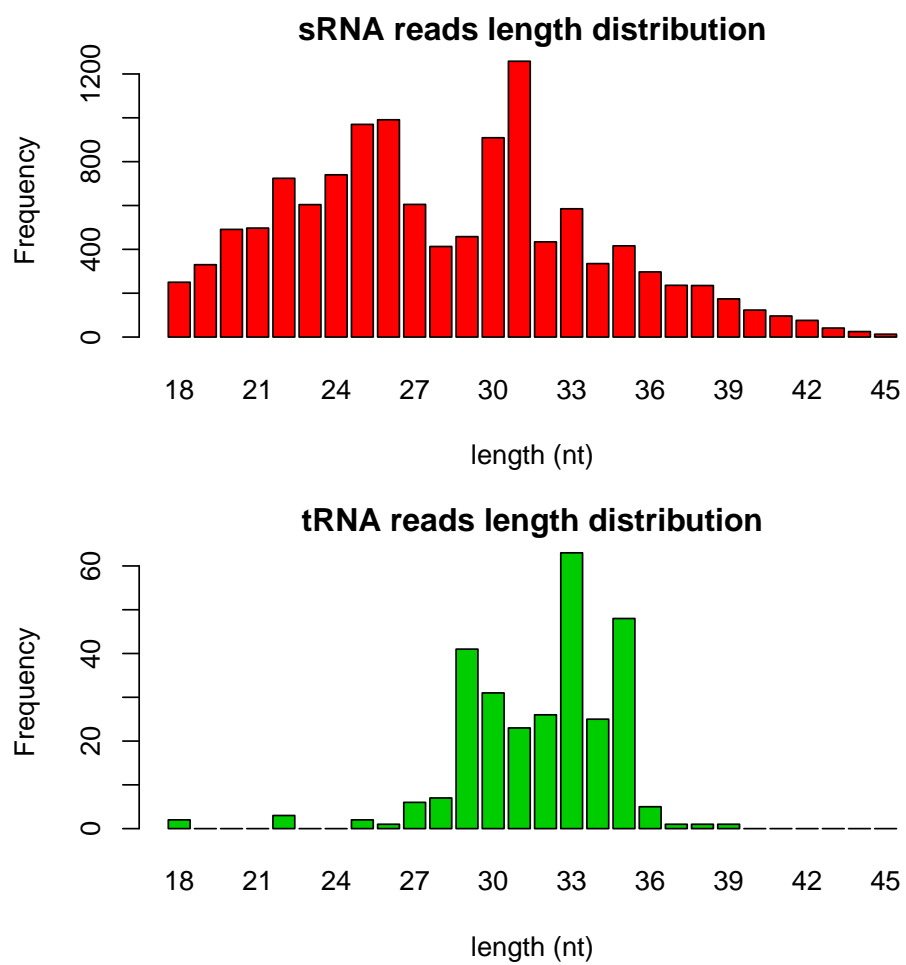


Figure 4: Small RNA and tRNA reads distribution

the read.

```
sRNA example:  
>Abc0000001_772  
CAGGTGGTCAGGTAGAGAATACCAAGGCGCT  
>Abc0000002_475  
AGGTGGTCAGGTAGAGAATACCAAGGCGCT
```

5. I have problem in installing tsRFinder and/or the dependencies, where to get more help?

You can go to the official support website or create new issue for tsRFinder repository on GitHub. The URL is <https://github.com/wangqinhu/tsRFinder/issues/new>

6. Where to report bugs?

Goto <https://github.com/wangqinhu/tsRFinder/issues/new>

7. Can we use tsRFinder for commercial purpose?

Yes. tsRFinder is free, open source software, see the MIT license.

8. The 'nwalgn' is not suit for my operating system, what can I do?

tsRFinder using Needleman-Wunsch algorithm nwalgn for small RNA alignment, we pre-build the binaries for some of the recently OS X / Linux, if you find it not suitable for your system or want to compile it by your self, goto lib/src directory and type 'make' to build ⁴ it.

⁴The gcc compiler is required.