Manual for tsRFinder

Qinhu Wang and Weixing Shan*

Northwest A&F University

Version 1.0.0

October 28, 2014

Abstract

The tsRFinder is a lightweight, fast and reliable tool for prediction and annotation of tRNA-derived small RNAs using next-generation sequencing data. It's a free open source software available at https://github.com/wangqinhu/tsRFinder.

*Email: wxshan@nwafu.edu.cn

Contents

1 Introduction

Small RNAs, such as miRNA, siRNA, and piRNA, are key regulators of gene expression. The tRNA-derived small RNA (tsRNA) is a recently identified novel class of small RNA, no public tool has tailored for tsRNA analysis yet. We thus developed tsRFinder for tsRNA prediction and annotation with additional sequence and statistical analysis, using small RNA sequencing data and the reference genome sequences.

2 How to install

2.1 Dependency

The tsRFinder depends on a few free open source softwares, please check and install them at first:

- Perl, v5.10.1 or higher, required for tsRFinder.pl execution. Always build-in in most of the UNIX-like OS.
  http://www.perl.org/get.html
- R, v2.15.2 or higher, required for small RNA data analysis and illustration.
  http://www.r-project.org
- bowtie, v1.0.0 or higher, required for small RNA mapping.
  https://github.com/BenLangmead/bowtie
- tRNAscan-SE, v1.3.1 or higher, required for tRNA prediction. It's optional if you prefer manual tRNA input.
  http://lowelab.ucsc.edu/tRNAscan-SE

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 may clone [1] tsRFinder by typing the following in the terminal.

Clone tsRFinder

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

Alternatively, you may download it from the following URL.

Latest release of tsRFinder

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

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

Setup tsRFinder

```
echo export tsR_dir="/the/path/of/tsRFinder" >> $HOME/.bashrc
source ~/.bashrc
```

tsRFinder is now ready for your dataset.

3 How to use

3.1 Preparing 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.

You are strongly recommended to use the reference genome sequence and raw small RNA sequencing data since tsRFinder is capable of preparing the reference tRNA data and clean small RNA data automatically. In case you prefer to preparing tRNA and small RNA files manually, see "demo/tRNA.fas" and "demo/sRNA.fa.gz" or access the format description in the FAQ section.

[1] If git is not installed, download it from http://git-scm.com

3.2 Running the pipeline

tsRFinder has two ways for arguments input, either configuration file or command line option. We recommend you to use a command line option for debugging and building your configuration file. Once your inputs are determined, you can write it to a configuration file for analysis.

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

Usage of tsRFinder: ./tsRFinder.pl -h

```
tsRFinder usage:

    tsRFinder.pl <option>

    -c  Configuration file
    -l  Label
    -g  Reference genomic sequence file, conflict with -t
    -t  Reference tRNA sequence file, conflict with -g
    -s  Small RNA sequence file
    -a  Adaptor sequence
    -n  Min read length            [default 18]
    -x  Max read length            [default 45]
    -e  Min expression level       [default 10]
    -u  Mature tsRNA level cut-off [default 10]
    -f  Small RNA family threshold [default 72]
    -w  tRNA with/without label    [default no/yes]
    -o  Output compressed tarball  [default no/yes]
    -m  Mode, run/debug            [default run/debug]
    -h  Help
    -v  Version

Examples:

    tsRFinder.pl -c demo/tsR.conf

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

4 Demo

To have a quick but rough overview that how tsRFinder looks like, see 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 sequences, in FASTA format and accessible from file "tsRFinder/demo/genome.fa". A reference genome sequence is also applicable. Figure 1 shows a refseq file in FASTA format.

Figure 1: Screenshot of the reference sequence in FASTA format

Demo sRNA: we extracted a bit of raw reads from some experimental data as a demo. Each read of the raw small RNA data have 4 lines. Beginning with "@" identifier line, followed by sequence line, "+" identifier optional line and quality score line, as shown in Figure 2.

4.2 Demo running

tsRFinder allows you specify inputs via a separate configuration file, as shown below for an example of the content of demo tsR.conf:

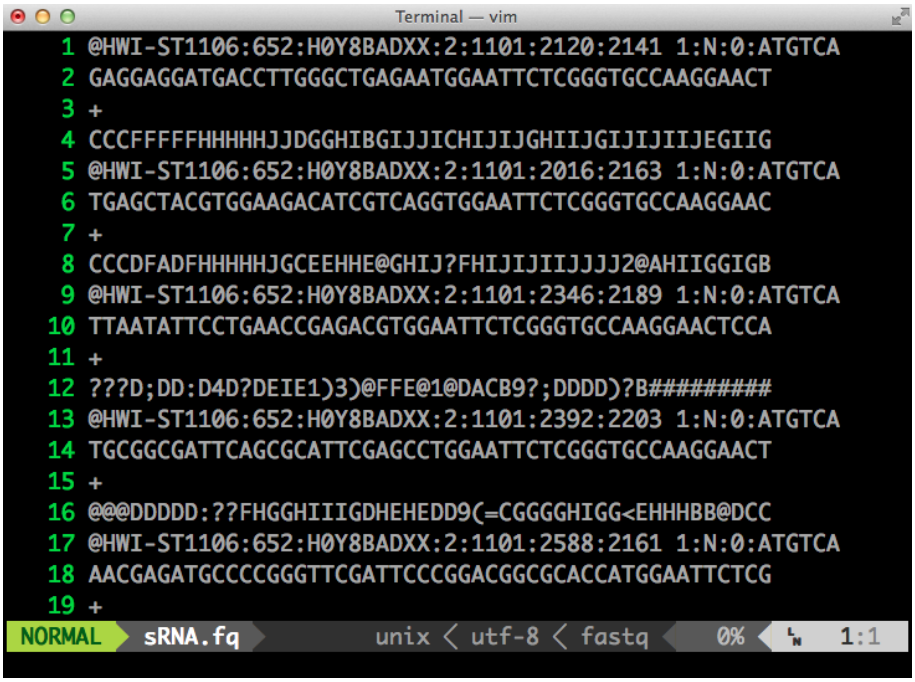

Figure 2: Screenshot of the small RNA sequence in FASTQ format

Demo configuration file for tsRFinder: demo/tsR.conf

```
label                :  Abc
reference_genome     :  demo/genome.fa
reference_tRNA       :
sRNA                 :  demo/sRNA.fq.gz
adaptor              :  TGGAATTCTCGGGTGCCAAGG
min_read_length      :  18
max_read_length      :  45
min_expression_level:  10
mature_cut_off       :  10
family_threshold     :  72
tRNA_with_label      :  no
output_compressed    :  no
```

Currently we have 12 items (16 options in command line) for configuration file filling. The argument items and the inputs are separated by colon (":"). You are recommended to use the first three letters of the investigated organism as a label (e.g. Ath for *Arabidopsis thaliana*, *-l* in command line

option); the paths of reference genome (*-g*) and small RNA (*-s*) should be supplied at least (tsRFinder supports input in gzipped files for sRNA and reference genome, and plain ASCII text is also acceptable). If raw sequence data is used, the adaptor sequence (*-a*) is required. In case a reference tRNA (*-t*) is not used, leave the argument to EMPTY.

Once the configuration file (-c) is prepared, an analysis protocol have been determined. In the demo, the configuration file is at demo/tsR.conf, typing the following in the terminal to run tsRFinder:

Running tsRFinder demo

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

To run alternative tRNA and sRNA dataset prepared yourself, typing the following:

Running tsRFinder alternative demo

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

tsRFinder allows you specify the minimum (*-n*) and maximum (*-x*) length of sequence for processing, you may specify the minimum expression level (*-e*) of the reads to increase reliability, as well as cut-off (*-u*) of the expression level of mature tsRNA. In case the family members are loose, the users may increase the tsRNA family threshold (*-f*) to tune it.

tsRFinder can work in run and debug modes (*-m*). In case of accessing some problems or some temporary files, you may enable the debug mode, otherwise use the run mode (default). To check the usage or version, you can use *-h* and *-v* option, respectively. To create a compressed tarball of the output file, use *-o yes* option please.

4.3 Demo output

By default, tsRFinder delivers a summary files with some basic statistics. The predicted or user applied tRNA sequence, the small RNA clean data, the tRNA reads, and the predicted tsRNA sequences are listed. A figure showing the length distributions of small RNA and tRNA reads (Figure 3) are included. Meanwhile, tsRFinder gives additional summary on tsRNA

family, and tRNA/tsRNA expression levels (including 5′ tsRNA, 3′ tsRNA and their abundance), text map (tmap) showing mapping of small RNA to tRNA, graphics showing the expression level based on small RNA data, the cleavage sites, and the cleavage profile (Figure 4). Shown below is a demo:

tsRFinder demo output list

```
---------
 SUMMARY
---------

*     tRNA seq : Abc/tRNA.fa
           Total : 5
*    sRNA reads : Abc/sRNA.fa
           Total : 10005215
          Unique : 7227
*    tRNA reads : Abc/tRNA.read.fa
           Total : 236122
          Unique : 50
*     tsRNA seq : Abc/tsRNA.seq
           Total : 7
          Unique : 6
* tsRNA report : Abc/tsRNA.report.xls
  tsR-5p total : 5
  tsR-3p total : 2
 tsR-5p unique : 4
 tsR-3p unique : 2
*      Text map : Abc/tsRNA.tmap
*    Visual map : Abc/images
* Distribution : Abc/distribution.pdf
*      Cleavage :
          Detail : Abc/cleavage.txt
         Profile : Abc/cleavage_profile.pdf
* tsRNA family : Abc/tsRNA.fam
* Stat. by BDI :
    Sensitivity : 0.9593
    Specificity : 0.7825
       Accuracy : 0.8819
```

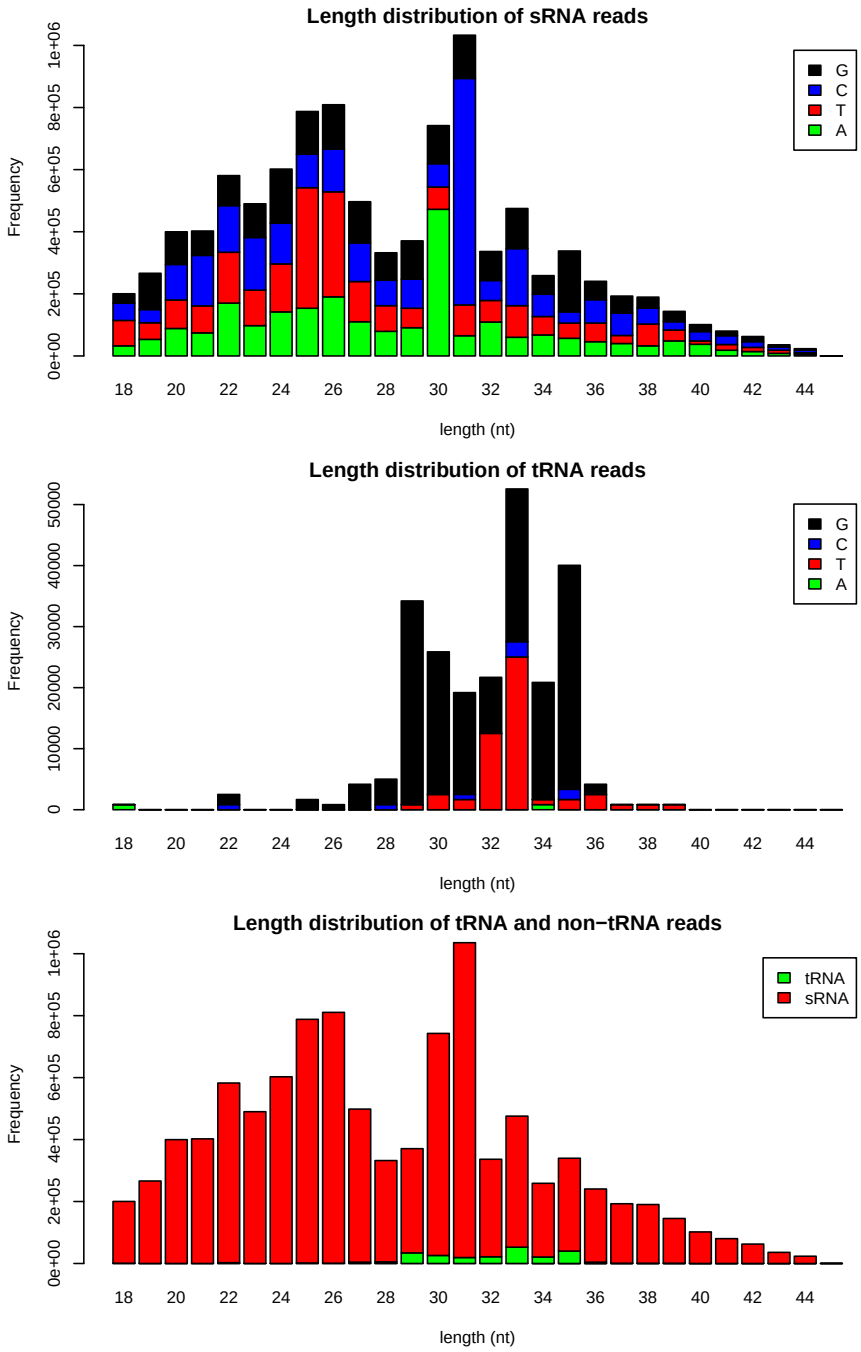

Figure 3: Length distribution of small RNA and tRNA reads

Note: The range of the small RNA length exhibited here is based on the minimum (*-n*) and maximum (*-x*) sequence length that specified in command line or the configuration file.

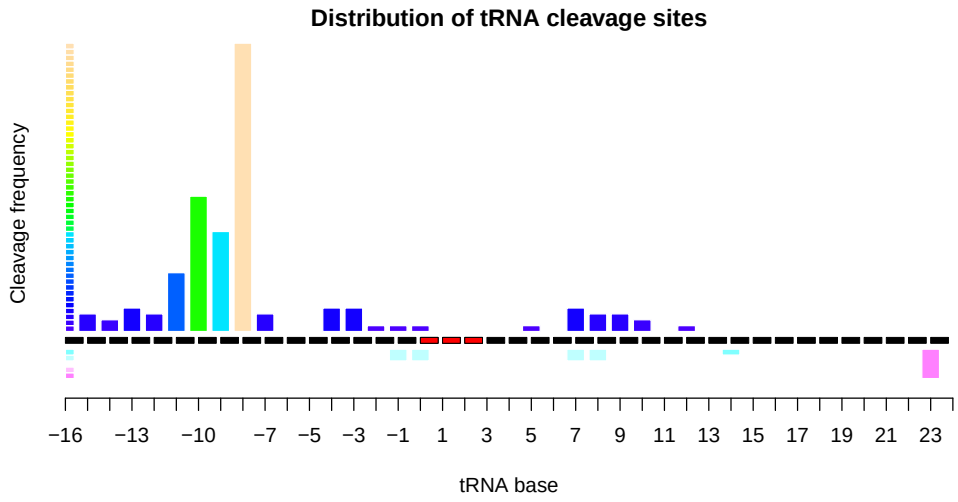

Figure 4: tRNA cleavage profile

Note: This figure is based on TAIR10 and *Arabidopsis thaliana* small RNA data GSM154336. The demo data contains only a few reads with limited cleavage information was not shown.

4.4 Visualization of tmap data

To examine the map of tsRNA, a vim syntax plugin was developed for visualization. You may enable color text map by copying lib/tmap.vim into the vim syntax folder, and put the following line into your .vimrc file:

Set filetype tmap in vim

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

Once tmap vim is installed, open the tsRNA.tmap file with vim to generate a color text map, as shown in Figure 5.

Visualization tsRNA.tmap

```
vim tsRNA.tmap
```

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

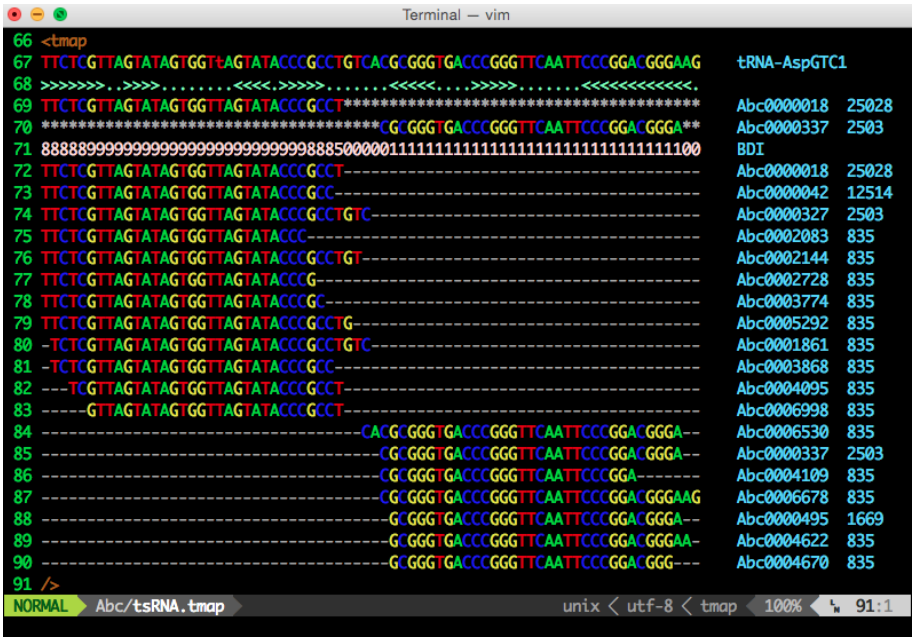

Figure 5: Screenshot of color tmap

5 FAQ

1. Can tsRFinder run on Windows?

No. tsRFinder depends on tRNAscan-SE (UNIX source code) and uses the advantages of some build-in program of UNIX-like systems, for example, awk, grep and head. Running on Windows may lead unexpected errors. Linux or OS X is strongly recommended.

2. What's the length required for small RNA reads?

Ideally, sequences from 15 - 50 nt for small RNA are recommend.

3. What's the tRNA format required for input?

You are encouraged to use reference genomic sequences since tsRFinder is capable of extracting tRNA sequences via tRNAscan-SE, removing duplicated sequences, and formatting the input. In case a manually prepared file is preferred, proceed following this format: first line, begin with ">", followed by label "Abc", and the name of tRNA "$tRNA-ProCGG1$"; second line, the sequence of tRNA; and third line, the secondary structure of tRNA, "><" and "()" both are acceptable. Be carefully in preparing this

because ">" equals "(" but not ")".
Shown below is an example.

tRNA example

```
>AbctRNA-ProCGG1
GGCCTCGTGGTCTAGTGGTATGATTCTC [NNN] AGAGGtCCCGGGTTCGATTCCCGGTGAGGCCC
>>>>>>>..>>>.........<<<.>>> [<.>] <<.....>>>>>.......<<<<<<<<<<<.
```

4. What's the sRNA format required for input?

You are recommended to use raw data. In case a manually prepared sRNA file is used for input, proceed following this format: first line, begin with ">", and the label "*Abc*", then the 7bit index, followed by "_" (or "|", "−" and white space) and the read number; second line, the sequence of the read. One more thing, the format of sRNA data in tsRFinder is compatible with the FASTX_TOOL kit [2], which helps you on processing the raw sequencing data.
Shown below is an example.

sRNA example

```
>Abc0000001_772
CAGGTGGTCAGGTAGAGAATACCAAGGCGCT
>Abc0000002_475
AGGTGGTCAGGTAGAGAATACCAAGGCGCT
```

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

You can access the official support website for trouble-shooting, or open 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 'nwalign' not compatible to my operating system, what can I do?

[2] http://hannonlab.cshl.edu/fastx_toolkit/

tsRFinder uses Needleman-Wunsch algorithm nwalign for small RNA alignment, with pre-building of the binaries for some of the recent OS X / Linux. If you find it not suitable for your system or want to compile it by yourself, goto lib/src directory and type 'make' to build [3] it manually.

[3]The gcc compiler is required. For OS X, you can install Xcode (ship with gcc)