# Manual for tsRFinder

Qinhu Wang and Weixing Shan*

Northwest A&F University

Version 0.9

August 8, 2014

**Abstract**

The tsRFinder is a lightweight, fast and reliable tool for transfer RNA-derived small RNA prediction and annotation with next-generating 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 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 specially tailored for tsRNA analysis yet. We thus developed tsRFinder for tsRNA prediction and annotation. 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 Dependency

The tsRFinder depends on a few free open source softwares, please check and install them [1] at first:

- Perl, greater than v5.10.1, required, for tsRFinder.pl execution.
  Always build-in in most of the UNIX-like OS.
- R, greater than v2.15.2, required, for small RNA data analysis and illustration.
  `http://www.r-project.org`
- bowtie, greater than v1.0.0, required, for small RNA mapping.
  `https://github.com/BenLangmead/bowtie`
- 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`

## 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

[1] The dependencies version were based on the oldest environment we have tried.

may save you a lot of time for trouble shooting.

First, you can clone [2] tsRFinder by typing:

Clone tsRFinder

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

in the terminal. Alternatively, you can download it from

Latest release of tsRFinder

```
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 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
```

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

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

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 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 using 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.

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

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 serval 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; if you don't have reference genome sequence, you can use the reference tRNA sequence instead. Figure 1 shows what a refseq file in FASTA format looks like.


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, Begin with "@" identifier line, followed by sequence line, "+" identifier optionally line and quality score line, just like what we have show in Figure 2.

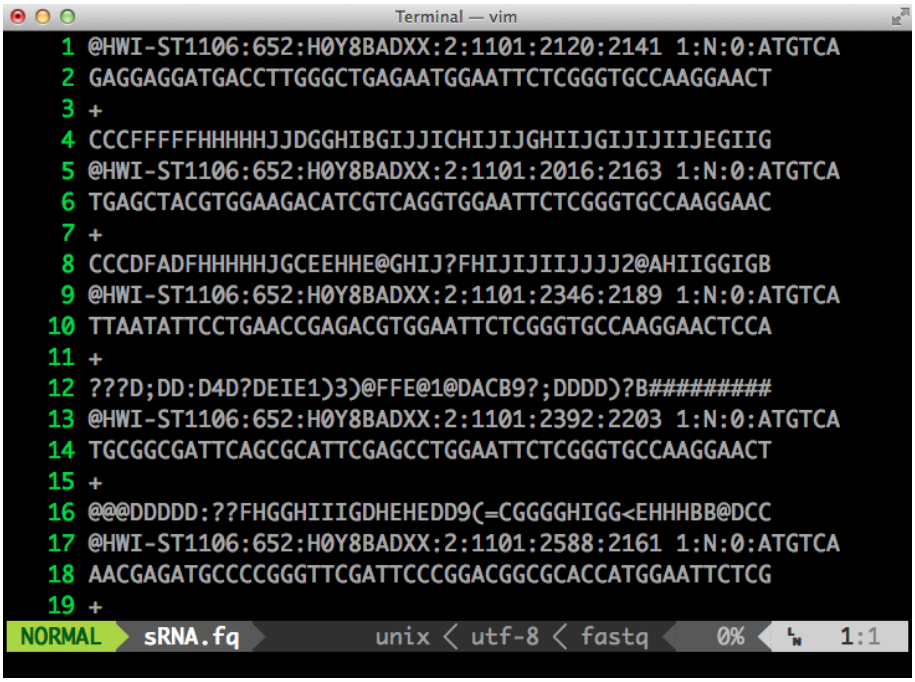


Figure 2: Screenshot of the small RNA sequence in FASTQ format

## 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:

Demo configuration file for tsRFinder: demo/tsR.conf

```
mode                 :  run
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 13 items (16 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 (tsRFinder support gzipped files input for sRNA and reference genome, plain ASCII text input is also acceptable). 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:

Running tsRFinder demo

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

tsRFinder allow you specify the minimum (*-n*) and maximum (*-x*) read length for processing, you can specify the minimum expression level (*-e*) of the reads to increase reliability, also the mature tsRNA expression level cut-off (*-u*). If you find the family member are too loose, increase the tsRNA family threshold (*-f*) to tune it.

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). 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 will give you a summary of which files have been outputted and some basic statistics. The predicted or user inputed tRNA sequence, the small RNA clean data, the tRNA reads, and the predicted tsRNA sequence were listed. A figure showing the small RNA and tRNA reads length distribution (Figure 3) were included. Meanwhile, tsRFinder gives you additional summary on tsRNA family, tRNA/tsRNA expression

(including 5′ tsRNA, 3′ tsRNA and their numbers), text map (tmap) of small RNA mapped to tRNA, graphics showing the expression evaluated by small RNA data, the cleavage site, and also the cleavage profile (Figure 4). See our demo summary here:

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-5p 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
```

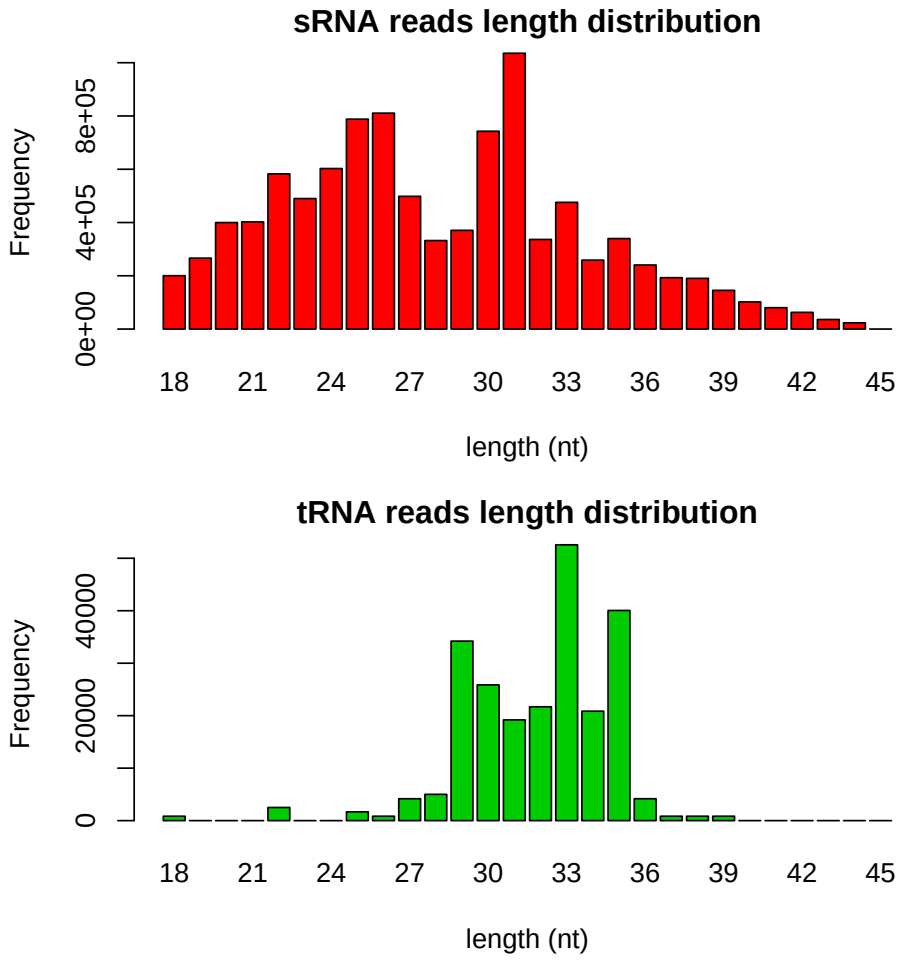


Figure 3: Small RNA and tRNA reads distribution

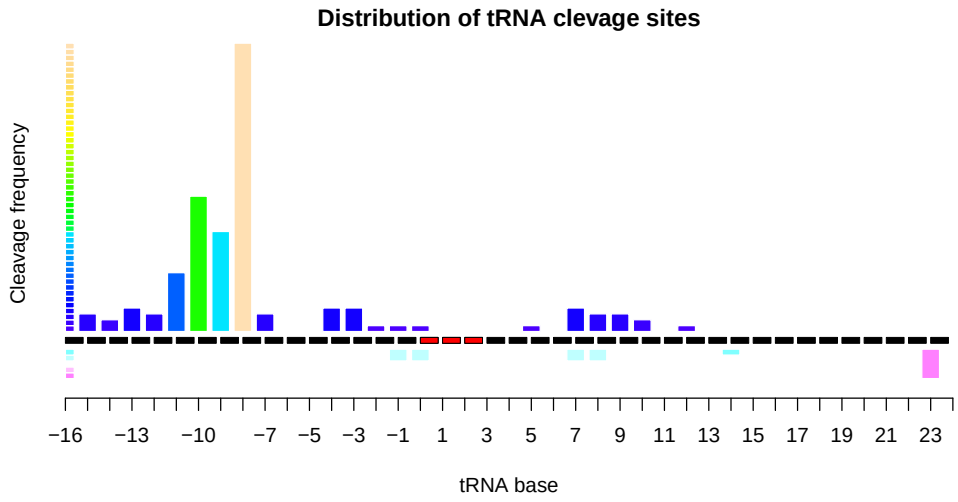


Figure 4: tRNA cleavage profile [3]

### 4.4 Visualization of tmap data

To examine the map of tsRNA, we developed a vim syntax plugin for visualization. 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:

Set filetype tmap in vim

```
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 Figure 5.

Visualization tsRNA.tmap

```
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?**

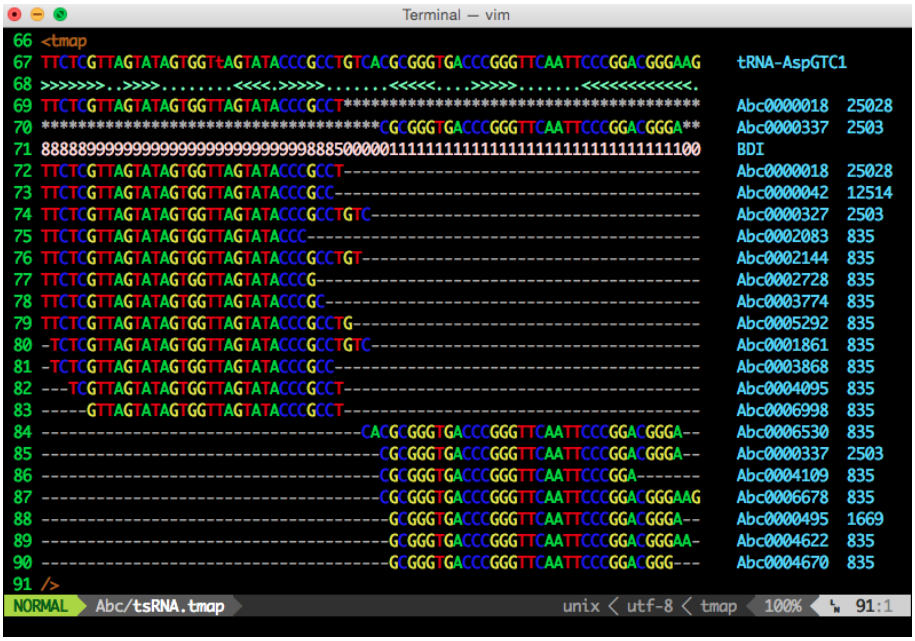


Figure 5: Screenshot of color tmap

No. tsRFinder depends on tRNAScan-SE (UNIX source code) and use the advantages of 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.

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

Ideally, we recommend you sequencing from 15 - 50 nt for small RNA. However, 18 - 30 nt is OK if your tsRNA is shorter than 30 nt (such as tRF).

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

At first we encourage you using a reference genomic sequence to instead it, since tsRFinder will help you extracting the tRNA sequence via tRNAscan-SE, removing duplicated sequences, and formatting input. And if you really want to use a file prepared yourself, please follow this format (see the example): first line, begin with ">", followed by label "$Abc$", then which tRNA "$tRNA - ProCGG1$"; second line, the sequence of the tRNA; and third line, the secondary structure of tRNA, "><" and "()" both were acceptable. However, be carefully in preparing this because ">" equals "(" but not ")".

tRNA example

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

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

We recommend you using a raw data. To have an sRNA file for input prepared yourself, please follow this format (see example): first line, begin with ">", then 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, tsRFinder sRNA data format is compatible with FASTX_TOOL kit [4], so you can prepare this with fastx tools.

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 'nwalign' is not suit for my operating system, what can I do?**

tsRFinder using Needleman-Wunsch algorithm nwalign 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 [5] it.

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

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