Manual for tsRFinder

Qinhu Wang and Weixing Shan*

Northwest A&F University

July 12, 2014

Contents

*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 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 prediction yet. We thus developed tsRFinder for tsRNA prediction. It takes the raw data of small RNA sequencing reads 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 the following programmes, please check and install them [1] at first:

- Perl, greater than v5.10.1, required, for tsRFinder.pl execution. And 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 havn'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 yourself.
  https://github.com/agordon/fastx_toolkit

[1] The dependencies version is based on the oldest test enviroment we have.

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 clone [2] tsRFinder by typing:

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

in the terminal. Alternatively, you can download it from

```
https://codeload.github.com/wangqinhu/tsRFinder/zip/master
```

and then unzip master file.

When tsRFinder is cloned or unpacked, move the entire directory to an proper place and add the tsRFinder path to the 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 find out which exact format of tRNA reference and small RNA reads file

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

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
    -x  Max read length
    -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 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.

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.

4.2 Demo running

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

```
label               :  Abc
reference_genome    :  demo/genome.fa
reference_tRNA      :
```

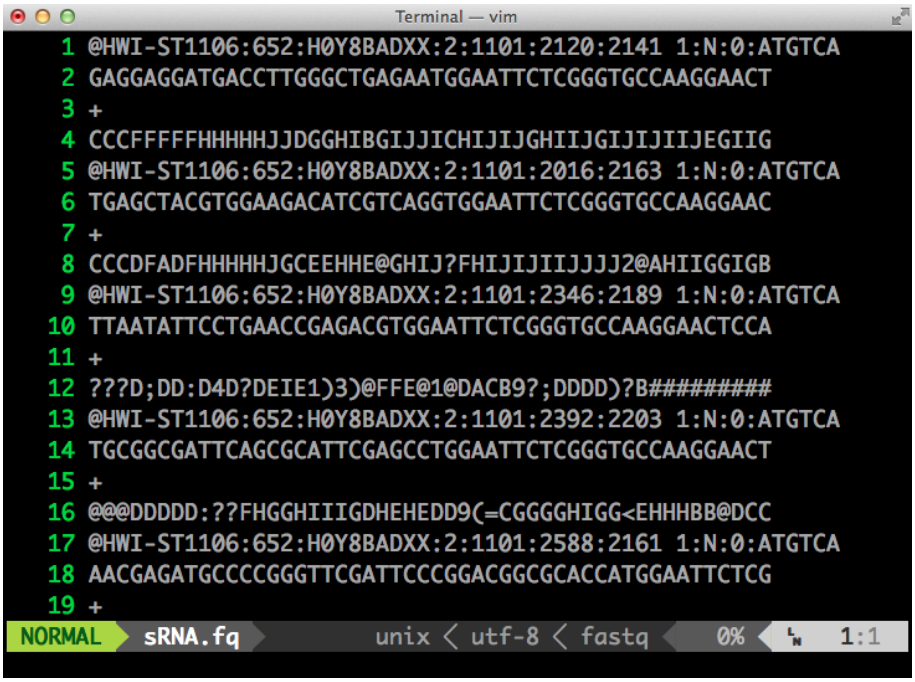

Figure 2: Screenshot of the small RNA sequence in fastq format

```
sRNA                :   demo/sRNA.fq
adaptor             :   TGGAATTCTCGGGTGCCAAGG
min_read_length     :   18
max_read_length     :   45
```

Currently we have 8 arguments used for 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
```

4.3 Demo output

By default, tsRFinder will give you a summary of which files have been outputted and some basic statistic. The predicted or user inputed 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, and also the cleavage site. Additional, tsRFinder supplies a figure showing the small RNA and tRNA reads length distribution (Figure 3) 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 : /Users/wangqinhu/tsRFinder/Abc/cleavage.txt

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

stat. by BDI :
 Sensitivity : 0.94
 Specificity : 0.7838
    Accuracy : 0.8764

---------
```

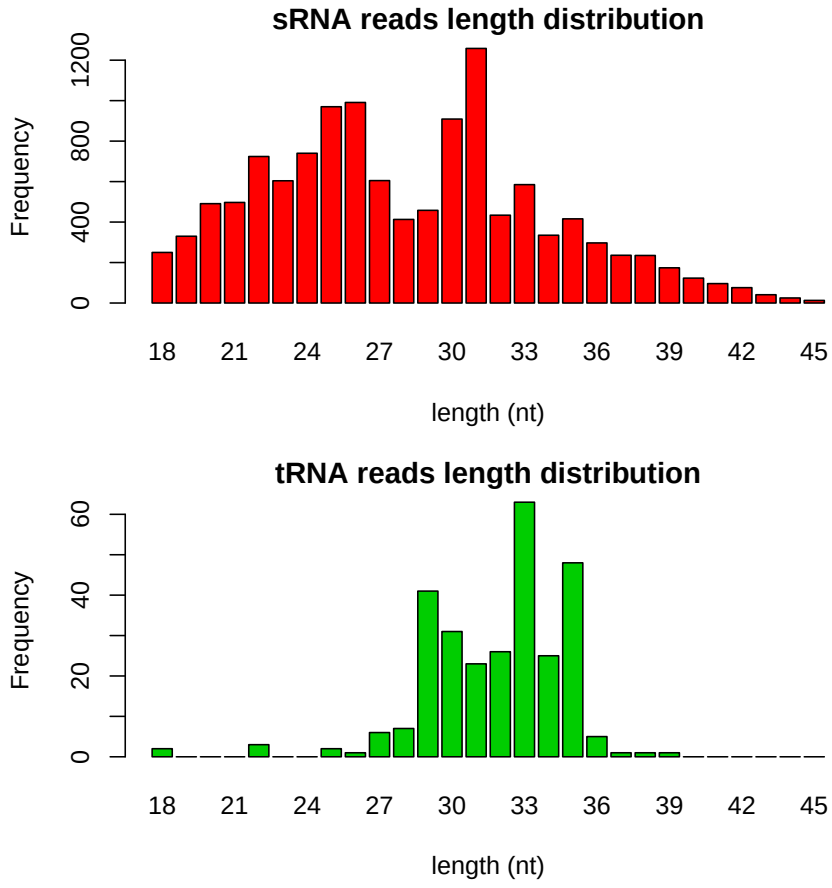

Figure 3: Small RNA and tRNA reads distribution

4.4 Visualization of tmap data

To examine the map of tsRNA, we developed a vim syntax plugin for visualization (Figure 4). 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
```

Figure 4: Screenshot of color 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.

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

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 extract the tRNA sequence via tRNAscan-SE. And if you really want to use a file prepared yourself, please follow this format (see 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.

```
tRNA example:
>AbctRNA-ProCGG1
GGCCTCGTGGTCTAGTGGTATGATTCTCGCTTCGGGTGCGAGAGGtCCCGGGTTCGATTCCCGGTGAGGCCC
>>>>>>>..>>>.........<<<.>>>>>.......<<<<<.....>>>>>.......<<<<<<<<<<<.
```

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.

```
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 offical 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 algorithorim 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 complile it by your self, goto lib/src directory and type 'make' to build [3] it.

[3]The gcc compiler is required.