Manual for tsRFinder

Qinhu Wang and Weixing Shan*

Northwest A&F University

July 3, 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), 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 identify tsRNA 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 v5.16.2, required, for tsRFinder.pl execution. And it was always installed already in most of the UNIX-like operating systems.
- tRNAscan-SE, greater than v1.3.1, required, for tRNA prediction.
  http://lowelab.ucsc.edu/tRNAscan-SE
- 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
- 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 dependency versions were 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. 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 PATH="/your/path/of/tsRFinder:$PATH" >> ~/.bashrc
echo export tsR_dir="/your/path/of/tsRFinder" >> ~/.bashrc
source ~/.bashrc
```

And now, you can running 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

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

the demo data and then find what exact format of tRNA reference and small RNA reads file can be accepted instead, this is allowed but not encouraged.

3.2 Running the pipeline

tsRFinder supplies two ways for inputs, 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

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 look 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 our demo tsR.conf:

```
label                 :   Abc
reference_genome      :   demo/genome.fa
reference_tRNA        :   NA
sRNA                  :   demo/sRNA.fq
adaptor               :   TGGAATTCTCGGGTGCCAAGG
min_read_length       :   18
max_read_length       :   45
```

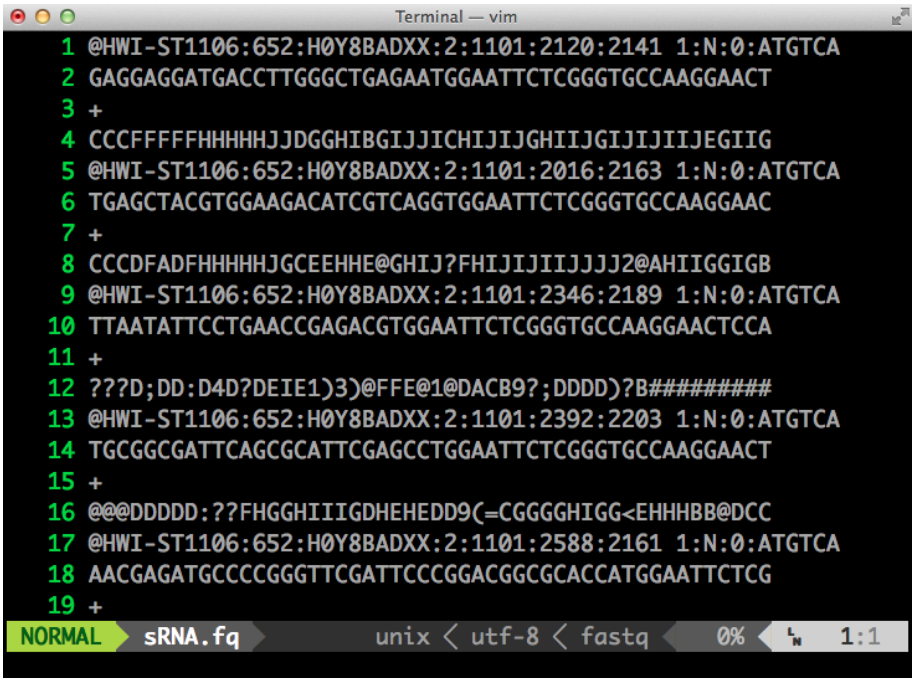

Figure 2: Screenshot of the small RNA sequence in fastq format

Currently we have 7 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*); the paths of reference genome and small RNA should be also supplied at least. If you are using a raw sequence data, please also input the adaptor sequence, and the shortest read you want to keep. If you are not using a reference tRNA prepared by yourself, leave this argument to "NA" please.

Once your configuration file is well prepared, typing the following in the terminal to run tsRFinder. In this demo, the configuration file 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), the text map (tmap) of small RNA mapped to tRNA, the graphics showing the expression evaluated by small RNA data, 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

stat. by BDI :
 Sensitivity : 0.940025252525252
 Specificity : 0.783809523809524
    Accuracy : 0.876447574334898

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

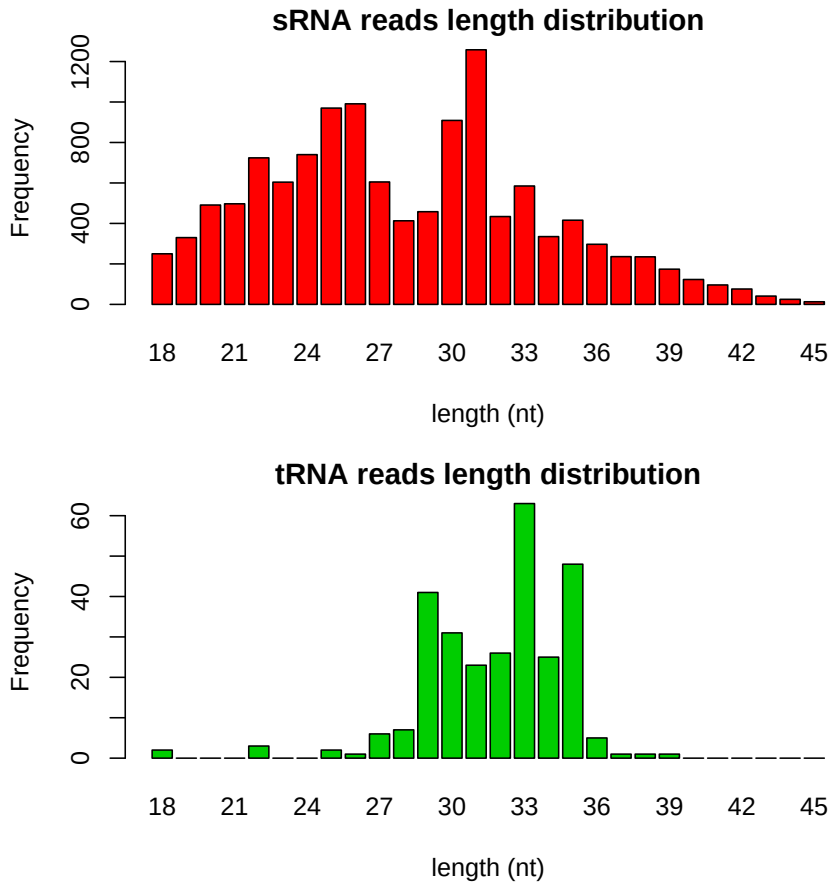

Figure 3: Small RNA and tRNA reads distribution

4.4 Visualization of tmap data

To examine the map the 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 running 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 less than 30 nt (such as tRF).

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

You can create new issue for tsRFinder repository on GitHub. The URL is https://github.com/wangqinhu/tsRFinder/issues/new

4. Where to report bugs?

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

5. Can we use tsRFinder for commercial purpose?

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