zhoujj2013 / **Tag-seq**

Data analysis pipeline for Tag-seq.

17 commits | 2 branches | 0 packages | 0 releases | 1 contributor | MIT

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

zhoujj2013 Update README.md — Latest commit aae1ad0 12 minutes ago

bin	v1.1	11 hours ago
test	v1.1	11 hours ago
.gitignore	Create .gitignore	11 hours ago
LICENSE	Create LICENSE	17 hours ago
README.md	Update README.md	12 minutes ago
global.png	v1.1	11 hours ago
offtargets.png	v1.1	11 hours ago
python.package.requirement.txt	v1.1	11 hours ago
sites.png	v1.1	11 hours ago
stat.txt	v1.1	11 hours ago

README.md

Tag-seq

Data analysis pipeline for Tag-seq.

System requirements

Tag-seq runs under the Linux (i.e., Centos, see also https://www.centos.org/ for further details) on a 64-bit machine with at least 32 GB RAM.

Tag-seq requires PERL v5, R, Python 2.7, pip and several python packages listed in python.package.requirement.txt.;

Tag-seq also requires some third-party packages:

STAR aligner
FASTQC
AdapterRemoval
BEDTOOLS
SAMTOOLS
PICARD
umi_tools
bedops
water in EMBOSS
RIdeogram

Tag-seq have been tested in CentOS release 7.4 (Linux OS 64 bit).

Installation

Get Tag-seq pipeline

```
git clone https://github.com/zhoujj2013/Tag-seq.git --depth 1
```

Preparation

Download reference genome and build index.

```
# download genome
mkdir hg19
cd hg19
wget http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz
wget http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/hg19.chrom.sizes
gunzip hg19.fa.gz

# build genome index
/path_to/STAR  --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles ./hg19.fa --runThreadN 32
```

Run test

If you have obtained the reference genome, STAR index, you can run test to examine whether the package works well (the test dataset is placed in ./test directory within Tag-seq).

Tag-seq requires a configure file containing paths of input files, sgRNA, Tag primers and genome etc. (See config.TEST.txt for more details.)

```
cd test

# run
sh work.sh

# around 30 mins.
# you can check the report in out.XXX/find.target/.
```

Result

1. QC statistics

you can check stat.txt.

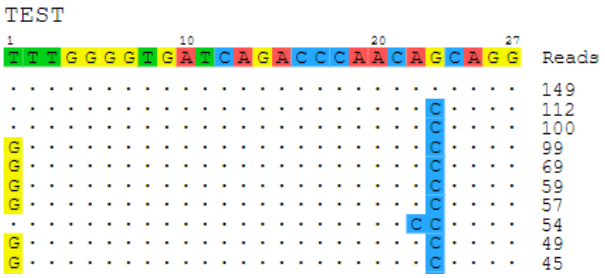
2. Information of potential targets in bed format

```
chr1    10111   10112   AAVS1.E_minus_minus_2_9,AAVS1.E_plus_minus_1_13 0       29      0       12
chr1    55903742        55903743        AAVS1.E_minus_minus_3669_8,AAVS1.E_plus_minus_5324_6    0       11      0
17
chr1    68164302        68164303        AAVS1.E_minus_minus_4802_6,AAVS1.E_plus_plus_6944_6     8       0       0
7
chr1    111700139       111700140       AAVS1.E_minus_minus_7763_6,AAVS1.E_plus_plus_11377_6    9       0       0
5
chr1    121478642       121478643       AAVS1.E_minus_plus_8420_6,AAVS1.E_plus_plus_12435_6     9       0       7
0
```

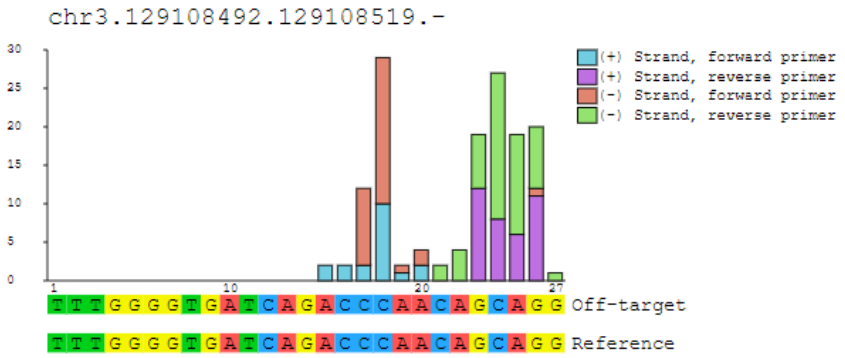
Column 1: chromosome
Column 2: start
Column 3: end
Column 4: id
Column 5: read count for plus strand in plus library
Column 6: read count for minus strand in plus library
Column 7: read count for plus strand in minus library
Column 8: read count for minus strand in minus library

3. Potential off-targets

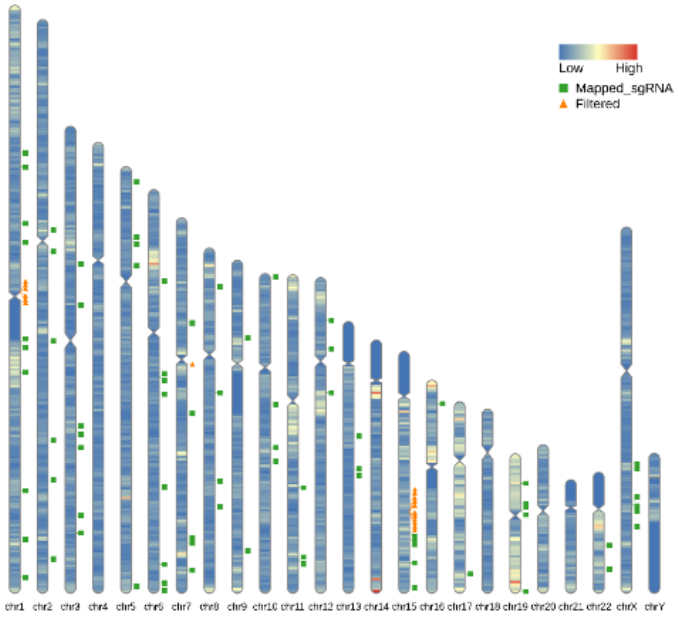
Illustrate of off-targets sites and read count.



4. Read counts across sgRNA in target and off-target sites



5. Global view of target and off-target sites



Tag-seq Runtime

The running time of Tag-seq depends on the size of sequencing depth (For 30M flagments, it takes 30mins).

Please cite

1. xxxx Tag-seq (underreview)