scFusion DOI 10.5281/zenodo.6450470

scFusion is a computational pipeline for detecting gene fusions at single-cell resolution. scFusion works on Linux/Mac OS. If you have any questions related to scFusion, please visit https://github.com/ZijieJin/scFusion and post them on the *Issues* page or email me: jinzijie@pku.edu.cn

Software Prerequisite

The software below should be in your PATH. **(They can be installed by conda and pip)**

- STAR >= 2.7.2d (tested on 2.7.2d and 2.7.8a)
- samtools >= 1.0 (tested on version 1.10)
- bedtools (tested on version 2.29.2)
- python 3
- R >= 3.5
- R package: stringr
- python module: pyensembl
- python module: pysam (tested on version 0.18.0)
- python module: tensorflow, keras, and numpy (version 2.8.0, 2.8.0, and 1.22.3, respectively) **OR** tensorflow, keras, numpy, and scipy (version 2.3.0, 2.4.3, 1.18.5, and 1.4.1, respectively)

Recommend Configuration

- 64 GB memory or more for each task
- 8 CPU cores or more for each task

Optional Configuration

- Job scheduler (e.g. Slurm)

Data Requirement

- Single cell RNA sequencing files from Smart-Seq protocol. File names should be **1.fastq,* *2.fastq (e.g. 1_1.fastq, 1_2.fastq, 2_1.fastq, 2_2.fastq) ** *_1.fq is not allowed.**
- Reference genome file (*.fa)(like hg19.fa, file size = ~3G)
- GTF annotation file (*.gtf) (Can be obtained from Ensembl (ftp://ftp.ensembl.org/pub/), NCBI, or UCSC)
- (Optional) Mappabilityfile (Can be obtained from UCSC) **If you are using the hg38 version of mappability file, please ensure the file format is the same as hg19's. (Only 4 columns. chr, start, end, value)** If you do not provide this file, scFusion will turn off the mappability filter.

Usage

Download all scripts, unzip the hg19mappability file in the data folder (if you want to use this mappability file), and run scFusion.py.

scFusion has 11 subcommands: BuildSTARIndex, Index, Rename, ReadMapping, ReadProcessing, FusionCandidate, Retrain, ArtifactScoring, FusionReport, RestoreName, DeleteTempFiles. You should run these commands in correct order (see the testdata section). (Or skip some commands)

You can type `python scFusion.py` or `python scFusion.py [subcommands]` to see the help.

BuildSTARIndex

This command builds the STAR index. If you already have it, you can skip this step. This step may cost 30 - 60 mins.

Required parameters are:

-s: specify the output folder of STAR index.

-g: specify the reference file (*.fa).

-a: specify the annotation file (*.gtf).

Optional parameters are (default values are in brackets):

-t: specify the number of threads scFusion can use [8].

Rename

This command renames files using consecutive numbers. scFusion only accepts input files whose names are numbers (but do not need consecutive numbers).

Required parameters are:

-f: specify the folder of input sequencing file.

Index

This command generates some necessary files and save them in GENOMEDIR. This step costs less than 3 mins.

Required parameters are:

-d: specify the path of GENOMEDIR.

-g: specify the reference file (*.fa).

-a: specify the annotation file (*.gtf).

ReadMapping

This command maps pair-end reads using STAR. We recommend you splitting the mapping task and using at least 12 threads for each mapping task. For example, if you have 1000 single-cell sequencing files, you can generate 100 mapping tasks and each task only contains 10 single cells.

Required parameters are:

-f: specify the folder of input sequencing files.

-b: specify the first index of files.

-e: specify the last index of files.

-s: specify the STAR index folder.

-o: specify the output folder of scFusion

Optional parameters are (default values are in brackets):

-t: specify the number of threads scFusion can use [8].

ReadProcessing

This command processes the chimeric reads reported by STAR. This step costs 3 - 5 mins for each single cell. Only ONE thread will be used in this step. We recommend you splitting the task into smaller tasks.

Required parameters are:

-d: specify the path of GENOMEDIR.

-b: specify the first index of files.

-e: specify the last index of files.

-o: specify the output folder of scFusion. It should be the same as that in read mapping step.

Optional parameters are (default values are in brackets):

-m: specify the mappability file. [scFusion/data/hg19mappability75.txt].

-M: To keep all reads and do not apply mappability filter [False].

FusionCandidate

This command identifies the fusion candidates. This step may cost several hours depending on the size of input data. Only ONE thread will be used in this step.

Required parameters are:

-d: specify the path of GENOMEDIR.

-b: specify the first index of files.

-e: specify the last index of files.

-o: specify the output folder of scFusion. It should be the same as before.

Optional parameters are (default values are in brackets):

-p: specify the prefix of fusion candidate files []. It should only be specified if users want to compare the results of different settings.

Retrain

This command retrains the network using input data. This step may cost one hour to several hours depending on the size of input data and the performance of computer. All available threads will be used in this step.

Required parameters are:

-o: specify the output folder of scFusion. It should be the same as before.

Optional parameters are (default values are in brackets):

-p: specify the prefix of fusion candidate files []. It should only be specified if users want to compare the results of different settings.

-w: specify the initial weight file of the deep-learning network [scFusion/data/weight-V9-2.hdf5].

-c: specify the number of epochs in the retraining step [10].

ArtifactScoring

This command detects artifacts using the network. This step may cost one hour to several hours depending on the size of input data and the performance of computer. All available threads will be used in this step.

Required parameters are:

-d: specify the path of GENOMEDIR.

-b: specify the first index of files.

-e: specify the last index of files.

-o: specify the output folder of scFusion. It should be the same as before.

Optional parameters are (default values are in brackets):

-p: specify the prefix of fusion candidate files []. It should only be specified if users want to compare the results of different settings.

-w: specify the weight file of the deep-learning network [scFusion/data/weight-V9-2.hdf5]. If you have retrained the network, please specify the retrained weight file (the path is provided when finish the retrain).

FusionReport

This command reports gene fusions detection by scFusion. This step may cost half an hour to several hours depending on the size of input data and the performance of computer. Only one thread will be used in this step.

Required parameters are:

-f: specify the folder of input sequencing files.

-b: specify the first index of files.

-e: specify the last index of files.

-d: specify the path of GENOMEDIR.

-o: specify the output folder of scFusion. It should be the same as before.

Optional parameters are (default values are in brackets):

-p: specify the prefix of fusion candidate files []. It should only be specified if users want to compare the results of different settings.

-v: specify the p-value(FDR) cutoff of the statistical model [0.05].

-n: specify the artifact score cutoff [0.75].

--LncRNAFilterOff: turn off the LncRNA filter.

--NoApprovedSymbolFilterOff: turn off the no-approved-symbol filter.

RestoreName

This command restores the file names using RenameList.txt in input data folder.

Required parameters are:

-f: specify the folder of input sequencing file.

DeleteTempFiles

This command deletes temporary files. scFusion cannot undo the deletion.

Required parameters are:

-o: specify the output folder of scFusion. It should be the same as before.

Optional parameters are:

--AllTempFiles: delete all temporary files generated by scFusion and STAR.

--AllUnimportantTempFiles: delete all unimportant temporary files generated by scFusion.

--STARMappingFiles: delete STAR mapping results.

Run the test data

```
python scFusion.py BuildSTARIndex -g hg19.fa -a ref_annot.gtf -s hg19STARIndex/ -t 20
```

```
python scFusion.py Index -g hg19.fa -a ref_annot.gtf -d scFusionIndex/
```

```
python scFusion.py Rename -f testdata/
```

```
python scFusion.py ReadMapping -f testdata/ -b 1 -e 10 -s hg19STARIndex/ -o testdata_out/ -t 20
```

```
python scFusion.py Readprocessing -b 1 -e 10 -d scFusionIndex/ -o testdata_out/
```

```
python scFusion.py FusionCandidate -b 1 -e 10 -d scFusionIndex/ -o testdata_out/
```

```
python scFusion.py Retrain -o testdata_out/
```

```
python scFusion.py ArtifactScoring -b 1 -e 10 -d scFusionIndex/ -o testdata_out/ -w testdata_out/weights/RetrainWeight.hdf5
```

```
python scFusion.py FusionReport -f testdata/ -b 1 -e 10 -n 0.9 -d scFusionIndex/ -o testdata_out/
```

```
python scFusion.py RestoreName -f testdata/
```

```
python scFusion.py DeleteTempFiles -o testdata_out/ --AllUnimportantTempFiles --STARMappingFiles
```

After running the above command, scFusion is expected to report IGHJ5-IGHA1 fusion in testdata_out/Result.abridged.txt and testdata_out/Result.full.txt.