PULL User Manual

(build 201904)

1. Program description

PULL is the abbreviation of Plant Full-length lncRNA Identity. Which take advantages of sequencing for terminal signal to get a relatively complete transcript units collection. Additionally, long read sequencing data can also be added to form a high credible collection. After get the full-length transcriptome, we combined several popular tools to evaluate the coding potential in order to get the widespread non-coding RNAs, especially the long-noncoding RNAs (lncRNAs).

2. Work environment preparation

The whole pipline is worked on python3 environment, so we recommend a convenient tool called anaconda help you to build the work environment.

Step1: Downlaod anaconda

```
wget https://repo.anaconda.com/archive/Anaconda2-2018.12-Linux-x86_64.sh
sh Anaconda2-2018.12-Linux-x86_64.sh
```

Step2: Creat a new python3 work environment

```
conda create -n Py3_pull -c bioconda --file requirements_py3.txt python=3
```

in addition to the requirements satisfied by conda, here are also some softwares you are supposed to installed already.

ncbi-blast(v2.7.1+)

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.7.1/ncbi-blast-2.7.1+-x64-linux.tar.gz

stringtie(v1.3.4d)

http://ccb.jhu.edu/software/stringtie/dl/stringtie-1.3.4d.Linux_x86_64.tar.gz

R(v3.4.2)

https://mirrors.tuna.tsinghua.edu.cn/CRAN/src/base/R-3/R-3.4.2.tar.gz

CAGEr(v1.22.3)

http://bioconductor.org/packages/release/bioc/html/CAGEr.html

3. Installation of PULL

git clone https://github.com/chenyanj/PULL.git

Unpack the compressed file, then PULL is ready for use!

4. usage of PULL

Firstly you are supposed to activate the work environment for Pull:

source activate Py3_pul

Then use the absolute path of Pull to call the tool

Pull.py 1.0.0 -- Plant full length lncRNA identity

Usage: Pull.py [options]

Options:

-h	Show this screen.
-v	Show version.
-g GENO	Genome fasta file
-r REF	Genome annotation file in gtf format.
-c CAGE	A directory contain cage sample.ctss files
-p PAS	A directory contain pas.ctss files
-- ngs NGS_BED	A bed12 format file assembled by NGS sequencing data
--ngs_bam NGS_BAM	A directory contain all bam files for different samples.
--pb PACBIO	A bed12 format file assembled by Pacbio sequencing data.

Input Files needed

Reference files for target genome are obligatory, including sequence fasta file and annotation file in gtf format.

-g reference genome sequence in fasta format, we recommend you only keep chromosomal sequence except for other scaffold sequence.

-r reference genome annotation file in gtf format, the same as -g, only genes in chromosomal are supposed to be kept in order to save computation space.

-c A directory contain cage sample.ctss files, data for infomation of transcript start site(tss), such as CAGE-Seq. The input file is optional, if not provided, accurate 5'end will not be calculated. If provided, you are supposed to rename

fiels in this directory as sample.ctss .

-p A directory contain pas.ctss files, data for infomation of transcript end site(tes), such as PAS-Seq. The input file is optional, if not provided, accurate 3'end will not be calculated. If provided, you are supposed to rename fiels in this directory as sample.ctes .

--ngs A bed12 format file assembled by NGS sequencing data, you should assembel all sample to a final merged.bed. Raw sequencing data after quality contral are mapped to reference genome, then trascriptome file for every sample were merged together to get the final input bed12 file.

--ngs_bam A directory contain all bam files for different samples. In addition to the merged file, alignment results for every sample in bam format are also necessary for further dispose. The parameter accept a directory path in which bam files are named as sample.bam. Finally, the llibrary for NGS-Seq only be strand-specific library.

--pb A bed12 format file assembled by Pacbio sequencing data, The input file is optional. Raw sequencing data from pacbio is adjusted by NGS data, then fixed data is mapped to reference genome to get transcriptome file in bed12 format.

Provide tss or tes data should corresponding to ngs data, the sample name should be same. Ctss file are exampled bellow,

```
Chr01	27833	+	1
Chr01	28743	+	1
Chr01	28745	+	3
Chr01	28754	+	1
Chr01	28865	+	1
```

6.explanation for output files

This pipline will produce full-length lncRNAs collection in bed12 format which is classified by structural integrity to four types, including with-cap-and-polya, only-cap, only-polyA, other.