README

The codes included here are supplementary to our CPGAVAS2 web server, which can be accessed at http://www.herbalgenomics.org/cpgavas2. The code PREA.py can be used when the raw data size is too big to be uploaded to our CPGAVAS2 web server. The code for snp discovery and RNA editing site predction are for demonstration purpose only. It usage is beyond the scope of our work on CPGAVAS2 Web server. The users are expected to setup the appropiate environment by themselves uisng the demo code as an example.

CPGAVAS2 Team

Last updated: April 17, 2019.

==========

PREA.py

==========

PREA.py is a preprocessing tool intended to extract effective reads to reduce the size of NGS dataset for the prediction of RNA editing sites through CPGAVAS2 web server.

PREA.py has been tested on Linux system centos 6X, 7X, ubuntu 14, 16 and 18. PREA.py is developed using python v2.7 and biopython v1.68 and later. It should be pointed out that PREA.py does not support python v3 at this time. Software such as blast+, seqkit, seqtk have been provided with this script, please refer to license information of these software tools for their usage and cite the paper.

If you use this script, please cite our paper:
Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., & Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC genomics, 13(1), 715.

==========

Dependencies

==========

PREA.py requires the following environment:
1. Linux system
Centos 6X or later
2. python
python2.7, not support python3
3. BioPython
biopython 1.68 or later

==========

Usage

==========

python2.7 -q queryFile -t queryType[fasta|gb] -l leftFastqFile -r rightFastqFile -p outPrefix -m max_depth -o outdir

==========

Main arguments

==========

-h, --help	Print USAGE and DESCRIPTION; ignore all other parameters
-v, --version	Print version number; ignore other arguments
-q, --query	Input file name
-t, --type	Query file type, it should be exact “gb” or “fasta”
-l, --left	The left reads file in FASTQ format
-r, --right	The right reads file in FASTQ format
-p, --prefix	The prefix name for output file
-m, --max_depth	Maximum number of hits to keep
-o, --outdir	The output directory for depositing result files

The result files can be found in outdir with names: prefix_top[max?_depth]_1.fq and prefix_top[max?_depth]_2.fq

==========

Typical usage

==========

1. Select the reads for a particular set of genes.
   The target gene should be provided in FASTA format, the sample command line is as follow:

>python2.7 –q NC_000932.ndhB.fasta –t fasta –l reads_1.fq –r reads_2.fq –p Arth_ndhB -o outTmpDir –m 100

The result files are: “path/outTmpDir/Arth_ndhB_top100_1.fq” and
“ path/outTmpDir/Arth_ndhB_top100_2.fq”

2. For all genes from a full-length plastome sequence
   The annotation information should be provided in GenBank format. PREA.py will automatically extract the sequences of protein coding genes and CDS. The sample command line is as follow:

>python2.7 -q NC_000932.gb -t gb -l reads_1.fq -r reads_2.fq -p Arth_gene8cds -m 100 -o outTmpDir

The result files are: “path/outTmpDir/Arth_gene8cds_top100_1.fq” and
“ path/outTmpDir/Arth_gene8cds_top100_2.fq”

==========

Re-run with adjusting “--max_depth” argument.

==========

It is common that users might need to extract reads at different coverage. This can be achieved through the parameter " --max_depth".

When trying out different coverage, please keep the values of other argument unchanged and only modify the argument " --max_depth". This will avoid the step of re-running "makeblastdb" , which is usually the most time consuming step.

The output results can be also found in the output directory but with names with new max_depth like these: "path/outTmpDir/Arth_gene8cds_top[new_max_depth]_1.fq" and "path/outTmpDir/Arth_gene8cds_top[new_max_depth_2.fq".

==========END==========