PMiFish (Pipeline for MiFish data analysis)

Version 2.4.1 (2020/7/20)

1 . Summary

PMiFish is a pipeline written in Perl for analyzing eDNA data (mainly for MiFish data from Illumina sequencer). PMiFish uses Usearch software (Edgar, 2010) and Perl scripts written by author. The purpose of this pipeline is seamless analysis from start (fastq or gz file) to taxonomy assignment.

2. Configuration

＜for Windows users＞

・Install ActivePerl https://www.activestate.com/activeperl

・Install Gzip http://gnuwin32.sourceforge.net/packages/gzip.htm

After installing, user need to add manually "GnuWin32¥bin" to environment variables

・Download Usearch_v11 (https://www.drive5.com/usearch/download.html) and put it in Tools folder

＜for Mac users＞

・Download Usearch_v11 (https://www.drive5.com/usearch/download.html) and put it in Tools folder

*If you want to construct phylogenetic tree, you need to install MEGA X or more upper version (https://www.megasoftware.net/) , create a setting file (mao format) and put the file in Tools folder.

3. Folder description

DataBase folder … Put reference data file (fasta format) and Primer data file (txt format)

Dictionary folder … Put family and common name dictionary files

Results folder … Output files are generated in this folder

Run folder … Put fastq or gz files you want to analyze

Scripts folder … Perl scripts used for analysis are in this folder

Tools folder … Put Usearch program (and .mao file if you want to construct phylogenetic tree)

PMiFish.pl … Perl script

Setting.txt … Each parameter needs to write in this file

4. How to analyze

STEP1　Set each parameter in Setting.txt

STEP2　Put fastq or fastq.gz files generated by Illumina sequencer in Run folder

STEP3　Start PowerShell (windows)/Terminal (mac) at the PMiFish folder and input "perl pmifish.pl"

STEP4　Output files are generated in Results folder

5. Output files

*_Italic_ means Usearch option used

*If you set "Temporary = Yes" in Setting.txt, folders 1-1 to 1-4 were not generated

1 − 1　Merge_paird_reads　　*-fastq_mergepairs*

1 − 2　Strip_primers　　*-fastx_truncate* or Primer_Cleaner.pl

1 − 3　Quality_filter　　*-fastq_filter*

1 − 4　Rarefaction　　Rarefaction.pl

*If you set Rarefaction option in the Setting.txt, this folder was generated

2 − 1　Find_Unique　　*-fastx_uniques*

2 − 2　Denoise　　*-unoise3*

_unoise3_result.txt … Result files by denosing

2 − 3　Separate_chimera　　Separate_chimera.pl

_zotu_chimeras.fa … Chimeric sequences detected

_zotu_nonchimeras.fa … Sequences excluded chimeric sequence

2 − 4　Rarefaction　　Rarefaction.pl

*If you set Rarefaction option in the Setting.txt, this folder was generated

3 − 1　Usearch_global　　*-usearch_global*

4 − 1　Annotation　　Annotation.pl

_all_annotated_seq.fas … All unique sequences detected in sample

_Detail.html … Detailed results show this file

_Representative_seq.fas … Representative sequences

_Summary.txt … a summary file

_Synonym_list.html … If the same sequence detected from multiple species in reference data, this file was generated.

5 − 1　Fasta_for_Phylogenetic_Analysis

all_representative_seqs.fas … representative sequences from all samples

merged_list.txt … lists of species names that commonly detected among samples

merged_seq.fas … sequences combined the same sequences to one

merged_seq_with_family_name.fas … sequences with family name

5 − 2 Summary_table Summary_table.pl

Summary_table.tsv … A table summarizing the results in one

5 − 3 Fasta_classified_by_Family Fasta_classified_by_Family.pl

*If you set “Family = No” in Setting.txt, this folder was not generated

6 − 2 Phylogenetic_trees Phylogenetic_trees.pl

*If you set “Phylogenetic = No” in Setting.txt, this folder was not generated

log.html	The number of reads in each step
Portal.html	Portal.html summarize the results in each sample

[Author]
Ryo O. Gotoh
Natural History Museum and Institute, Chiba
rogotoh@chiba-muse.or.jp