1. The M-pick package includes a few files:
a. Programs to compute pairwise sequence distances: kmerdist_par and needledist (from ESPRIT [1])
b. Programs to do modularity clustering: community and convert (from Louvain [2])
c. Programs to run UCLUST as the preprocess to speed up: uclust.sh, otupipe.bash, and usearch (from OTUPIPE [3])
d. Files to link above programs:
Mpick.pl: the main OTU picking script;
MT_esprit_needle.pl: compute sequence distance using multiple threads;
gentree.pl: generate \epsilon-neighborhood graph;
genclusters.pl: generate the final clustering result;

2. To run M-pick
a. Unzip the files into a folder
b. chmod 755 *
c. In that folder run the perl script Mpick.pl.

```
Usage: perl Mpick.pl -s sequencefilename (with path) [-e epsilon]
[-d delta] [-f (0 or 1)]
Parameters:
    -s sequence input file, with relative or absolute path
    -e \epsilon, parameter to form a graph. Default 0.04
    -d \delta, parameter used for stopping criterion. Default 0.1
    -f 1:fast implement with uclust preprocessing; 0:regular
    implement. Default 0, time consuming but more accurate.
Output:
   clustering result in text format (*.mpick)
Example:
   perl Mpick.pl -s ../test/test.fas -f 1
#################################################################
```

d. A test sequence file (test.fas) is taken from [4] which contains simulated sequences from 11 taxa; each taxon has 10 sequences.
e. The last line of the generated *.mpick file holds the final clustering result.

3. Case study data
Data in Case studies 1 and 2 are uploaded. Unzip a zip file you will find 10 sequence files (*.fas) and corresponding annotation files (*.anotation). In a annotation file, a number in the last column represents the species information of a sequence.

Reference:
[1] Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, Farmerie W: ESPRIT: Estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* 2009, 37(10):e76.
[2] Blondel VD, Cuillaume JL, Lambiotte R, Lefebvre E: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics* 2008, P10008.
[3] Dgar RC, Haas BJ, Clemente JC, Quince C, Knight R: UCHIME improves sensitivity and speed of chimera detection, *Bioinformatics 2011*. doi: 10.1093/bioinformatics/btr381.

[4] Cheng L, AW Walker, J Corander: Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Res* 2012, doi:10.1093/nar/gks227.