Parallel-META 3 users' manual

Version 3.3
Release date: June/3/2016

Introduction

Parallel-META is a software toolkit which can perform rapid data mining among massive microbial community data for comparative taxonomical and functional analysis.

Based on parallel algorithms, Parallel-META can achieve a very high speed compare to traditional metagenomic analysis pipelines.

Parallel-META now supports 16S rRNA based taxonomical analysis and KEGG based predictive functional analysis for microbial community samples, phylogenetic and functional comparison and feature selection.

We strongly recommend that read this manually carefully before use Parallel-META.

Download

The latest release is available at:
http://www.computationalbioenergy.org/parallel-meta.html

Packages

Executive binary package:

Now the Parallel-META 3 executive binary package integrated with all tools is available for Linux (32 bit & 64 bit), which is easy to install.

Rscript environment:

For statistical analysis and pdf format output, Parallel-META 3 requires cran R (http://cran.r-project.org/) 3.0 or higher for the execution of ".R" scripts.

Source code package:

Parallel-META source code package is also available for building and installation for other Unix/Linux based operating systems.

Installation

Extract the package:

tar –xzvf parallel-meta-3.tar.gz

Configure the environment variables

export ParallelMETA=Path to Parallel-META
export PATH=”$PATH:$ParallelMETA/bin”
Rscript parallel-meta/Rscript/PM_Config.R

Compile the source code* (**only required** by the source code package):

cd parallel-meta
make

Package Dependency* (only required by source code package)

The following packages have been integrated in the executive binary package, but are still required by the source code package.

GCC 4.2 or higher;

Bowtie2 (2.1.0 or higher)

http://sourceforge.net/projects/bowtie-bio/files/bowtie2/
and put the “bowtie-align-s”to $ParallelMETA/Aligner/bin/

HMMER 3 :

http://hmmer.janelia.org/software/
and put the “hmmsearch” to $ParallelMETA/HMMER/bin/

Tools in toolkit

format-seq

For input sequence file(s) format check and reformat into Parallel-META required format of FASTA or FASTQ (See Sequence format). **We strongly recommend that check the sequence file(s) format before analysis steps.**

The **format-seq** accepts single input sequence file in FASTA or FASTQ format, or multiple input sequences in sample list format (See Sample list format). If the sequence passed the format check, f**ormat-seq** keeps the input with no modification; otherwise the **format-seq** reformats the input sequence file(s) with making a backup.

Usage:

format-seq [Options] Value

[Options]
-i Input sequence file in FASTA or FASTQ format [Conflict with -l]
-l Input sample list [Conflict with -i] (See Sample list format)
-p List file path prefix for '-l' [Optional]
-h help

Example:

format-seq –i sample.fasta

format-seq –l list.txt

pipeline

The pipeline is an integrated process for multiple samples with most of the following analysis steps.

Usage:

pipeline [Options] Value

[Option] :

Input options:

-i Sequence list file, pair-ended sequences are supported [Conflict with '-l'] (See **Sequence format and sequence list**)

-m Meta data file [Required] (See Meta-data format)

-l Taxonomic analysis results list [Conflict with '-i'] (See Sample list format)

-p List file path prefix for '-i' and '-l' [Optional]

Output options:

-o Output path, default is "default_out"

Profiling parameters:

-M Sequence type, T (Shotgun) or F (16S rRNA), default is F

-e Alignment mode, 0: very fast, 1: fast, 2: sensitive, 3: very-sensitive, default is 3

-P Pair-end sequence orientation, 0: Fwd & Rev, 1: Fwd & Fwd, 2: Rev & Fwd, default is 0

-r rRNA length threshold of rRNA extraction. 0 is disabled, default is 0

-k Sequence format check, T(rue) or F(alse), default is T

-f Functional analysis, T(rue) or F(alse), default is T

Statistic parameters:

-L Taxonomical levels (1-6: Phylum - Species). Multiple levels are accepted

-F Functional levels (Level 1, 2, 3 or 4 (KO number)). Multiple levels are accepted

-s Sequence number normalization depth, 0 is disabled, default is **disable**.

-b Bootstrap for sequence number normalization, default is 200, maximum is 1000

-R Rarefaction curve, T(rue) or F(alse), default is **F**

-E If the samples are paired, T(rue) or F(alse), default is F

-c Cluster number, default is 2
-T Network analysis edge threshold, default is 0.7

Other options:
-t cpu core number, default is auto
-h help

Notice:
1. Pair-ended sequences are supported for 16S rRNA sequences (See Sequence format and sequence list).
2. Samples (pairs) should be in the same order in all lists and meta-data.
3. In Meta data file, sample IDs should not be started with number, and should not contain space symbol (' ') and table symbol ('\t') (See Meta-data format).
4.Rarefaction curve is disabled in default setting, use "–R T" to enable (might be slow).
5. Sequence number normalization is disabled in default setting, use "-s" to enable and set the normalization depth (might drop samples with sequence number less than the setting depth).

parallel-meta

The **parallel-meta** accepts **single** shotgun sequences or 16S/18S rRNA sequences in FASTA or FASTQ format for taxonomical and predictive functional profiling.

Usage:

parallel-meta [Options] Value

[Options]
-m Input single sequence file (Shotgun)
A single input sequence file [Conflict with -r and -R]
-r Input single sequence file (rRNA targeted)
A single input sequence file [Conflict with -m]
-R Input paired sequence file for -r [Conflict with -m]
-o Output path
To appoint the result output path, default is "Result"
-e Alignment mode
0: very fast, 1: fast, 2: sensitive, 3: very-sensitive, default is 3
-P Pair-end sequence orientation for -R
0: Fwd & Rev, 1: Fwd & Fwd, 2: Rev & Fwd, default is 0
-k Sequence format check
T(rue) or F(alse), default is F
-L Integer value
To assign the rRNA length threshold of rRNA extraction. 0 is disabled, default is 0 [Conflict with -r]

-t Integer value

To assign the core number of CPU Parallel-META can use. Default is auto.

-f Functional analysis, T(rue) or F(alse), default is T

-h Help

Example:

parallel-meta –m examples/meta.fasta –o metaresults –e 1e-20 –l 150

extract-rna

The **extract-rna** can extract 16S & 18S rRNA fragments from shotgun sequences. This function has already been included in "parallel-meta" with parameter "-m" for shotgun sequences.

Usage:

extract-rna [Options] Value

[Options]

-m Input single sequence file (Shotgun)

To analysis a metagenomic shotgun file

-o Output Path

To appoint the result output path, default is "Extract_RNA"

-l integer value

To assign the 16S rRNA length threshold of RNA extraction, conflict with -r. 0 is disabled, default is 0.

-h Help.

Example:

extract-rna –m examples/meta.fasta –o metaresults –e 1e-20 –l 150

class-tax

For taxonomical analysis results integration of multiple samples. The **class-taxa** has already been integrated in program **parallel-meta.**

The **class-tax** accepts the taxonomical analysis results in "classification.txt" (see Results-Single sample).

Usage:

class-tax [Options] Value

[Options]

-i input filename [Conflict with –l]

-l Input filename list [Conflict with -i] (see Sample list format)

-p List file path prefix for '-l' [Optional]

-o output path, default is "./Result_Plot".

-h Help

Example:

class-tax –l list.txt –o result_plot

in which “list.txt” is the path of N samples’ taxonomical analysis results, and each sample in one line (see Sample list format).

class-func

For functional analysis of multiple samples. The **class-func** has already been integrated in program **parallel-meta**.

The **class-func** accepts the taxonomical analysis results in “classification.txt” (see Results-Single sample).

Usage:

class-func [Options] Value

[Options]
-i input filename [Conflict with –l]
-l Input filename list [Conflict with -i] (see Sample list format)
-p List file path prefix for '-l' [Optional]
-o output path, default is “./Result_Func”.
-t Cpu core number, default is auto
-h Help

Example:

class-func –l list.txt –o result_func

in which “list.txt” is the path of N samples’ taxonomical analysis results, and each sample in one line (see Sample list format).

class-func-nsti

For NSTI (Nearest Sequenced Taxon Index) value calculation of functional analysis. The **class-func-nsti** has already been integrated in program **parallel-meta**.

The **class-func-nsti** accepts the taxonomical analysis results in “classification.txt” (see Results-Single sample).

Usage:

class-func-nsti [Options] Value

[Options]
-i Input single filename [Conflict with -l and -T]
-l Input filename list [Conflict with -i and -T] (See Sample list format)
-p List file path prefix for '-l' [Optional]
-T Input OTU table format (*.Abd) [Conflict with -i and -T]
-o Output file, default is "NSTI.out"
-t Cpu core number, default is auto

-h Help

Example:

class-func-nsti –l list.txt –o result_func

in which “list.txt” is the path of N samples’ taxonomical analysis results, and each sample in one line (see Sample list format).

taxa-sel

For multi-sample feature selection (with a specified taxonomical level) based on the taxonomical profiling results.

The **taxa-sel** accepts the taxonomical analysis results in “classification.txt” (see Results-Single sample), and generate both the relative abundance table (*.Abd) and the absolute count table (*.Count).

Usage:

taxa-sel [Options] Value

[Options]
-l Input sample list [Required] (See Sample list format)
-p List file path prefix for '-l' [Optional]
-o Output file name, default is "taxaonomy_selection.txt"
-M Correlation matrix T or F, default is T
-L Taxonomical level (1-6: Phylum – Species, 7: OTU), default is 5
-r 16s rRNA copy number correction, T(rue) or F(alse), default is T
-p Print distribution barchart, T(rue) or F(alse), default is F
-q Minimum sequence count threshold, default is 2
-m Maximum abundance threshold, default is 0.001 (0.1%)
-n Minimum abundance threshold, default is 0.0 (0%)
-z Minimum No-Zero abundance threshold, default is 0.1 (10%)
-v Minimum average abundance threshold, default is 0.001 (0.1%)
-h help

Example:

taxa-sel –l list.txt –o taxa.txt –L 6

func-sel

For multi-sample feature selection (with a specified KEGG pathway level and relative abundance) based on the functional profiling results .

The **func-sel** accepts the functional analysis results in “functions.txt” (see Results-Single sample), and generate both the relative abundance table and the absolute count table.

Usage:

func-sel [Options] Value

[Options]
-l Input sample list [Required] (See Sample list format)
-p List file path prefix for '-l' [Optional]
-o Output file name, default is "taxaonomy_selection.txt"
-L KEGG Pathway level
Level 1, 2, 3 or 4 (4 is KO number) [Required]
-p Print distribution barchart, T(rue) or F(alse), default is F
-h Help

Example:
func-sel –l list.txt –o func.txt –L 2

comp-sam

For multi-sample comparison & similarity (distance) calculation based on the taxonomical profiling results.

The **comp-sam** accepts the taxonomical profiling results in "classification.txt" (see Results-Single sample).

Now from version 3.1 the **comp-sam** also accepts the relative abundance table (OTU level only) generated by **taxa-sel**. (see Results-Multiple Samples -Abundance_Tables)

Usage:

comp-sam [Options] Value

[Options]
-i Two samples path for single sample comparison [Conflict with –l and -T]
-l Sample name list table for multi-sample comparison [Confilct with – i and -T] (See Sample list format)
-p List file path prefix for '-l' [Optional]
-T Sample OTU table ((*.Abd results of **taxa-sel**) multi-sample comparison [Conflict with -i and -l]
-o Result output file, default is to output on screen
-r 16s rRNA copy number correction, T(rue) or F(alse), default is T
-d Output format, similarity (T) or distance (F), default is T
-p Print heatmap and clusters, T(rue) or F(alse), default is F
-t Cpu core number, default is 1
-h Help

Example:
Here suppose you compute the similarity of 2 samples:
comp-sam –i sample-1/classification.txt sample-2/ classification.txt –o sim.txt

or you have multiple samples:

comp-sam –l list.txt –o sim_matrix.txt –t 8

in which "list.txt" is the path of N samples' taxonomical analysis results, and each sample in one line (see Sample list format).

comp-sam-func

For multi-sample comparison & similarity (distance) calculation based on the functional profiling results.

The **comp-sam-func** accepts the functional profiling results in "functions.txt" (see Results-Single sample).

Now from version 3.1 the **comp-sam –func** also accepts the relative abundance table (KOlevel only) generated by **func-sel** (see Results-Multiple Samples -Abundance_Tables)

Usage:

comp-sam-func [Options] Value

[Options]

-i Two samples path for single sample comparison [Conflict with –l and -T]

-l Sample name list table for multi-sample comparison [Confilct with –i and -T] (See Sample list format)

-p List file path prefix for '-l' [Optional]

-T Sample KO table ((*.Abd results of **func-sel**) multi-sample comparison [Conflict with -i and -l]

-o Result output file, default is to output on screen

-d Output format, similarity (T) or distance (F), default is T

-p Print heatmap and clusters, T(rue) or F(alse), default is F

-t Cpu core number, default is 1

-h Help

Example:

Here suppose you compute the similarity of 2 samples:

comp-sam-func –i sample-1/functions.txt sample-2/ functions.txt –o sim.txt

or you have multiple samples:

comp-sam –l list.txt –o sim_matrix.txt –t 8

in which "list.txt" is the path of N samples' taxonomical analysis results, and each sample in one line (see Sample list format).

rare-curv

For rarefaction analysis and rarefaction curve printing to pdf format. It accepts the **absolute count table** (*.Count) results of **taxa-sel** and **func-sel**.

Usage:

rare-curv [Options] Value

[Options]

-i or –T Input feature table with Absolute Count (*.Abd results of **taxa-sel**) [Required]

-o Output file directory, default is "result"

-p Prefix name of output, default is "out"

-l Rarefaction curve label, T is enable and F is disable, default is F

-b The bootstrap value, default is 20

-t Cpu core number, default is auto

-h Help

Example:

rare-curv –i taxa.Count –o rare-out –b 20

comp-corr

For correlation calculation of taxonomical and functional distribution and with meta-data.

The **comp-corr** accepts the results of **taxa-sel** and **func-sel**.

Usage:

comp-corr [Options] Value

[Options]

-i Feature (Taxa/OTU/Function etc.) table file [Required]

-m Correlation file name

-o Output prefix, default is "corr_matrix"

-c Selected feature, separated by ","

-f S(Spearman) or P(Pearson) metrics,default is S

-N Network based co-occurrence analysis, T(rue) or F(alse), default is F

-T Netowrk analysis threshold, default is 0.7

-t Cpu core number, default is auto

-h Help

Example:

comp-corr –i taxa.txt –o taxa.network.txt

split-seq

For sequence split by barcode or group information

The **split-seq** accepts the input sequence in FASTA or FASTQ format.

Usage:

split-seq [Options] Value

[Options]
-i Input sequence file in FASTA or FASTQ format [Required]
-b Input barcode file [Conflicts with –g]
-g Input group file [Conflicts with –b]
-o Result output path, default is "Out"
-h Help

Notice: Sample IDs and sequence labels should not contain space symbol (‘ ‘) and table symbol (‘\t’). Here are examples of barcode file format and group file format:

The barcode file format:

ATTCGT	Sample1
AGCGTC	Sample2
……..	
CGTGAC	SampleN

The group file format:

Seq_Id_1	Sample1
Seq_Id_2	Sample2
……	
Seq_Id_N	SampleN

Example:

split-seq –i seq.fa –b barcode.txt –o seq.out
split-seq –i seq.fa –b seq.groups –o seq.out

split-table

For abundance table (*.Abd and *.Count) split by meta-data groups.
The **split-table** accepts the results of **taxa-sel** and **func-sel**.
Usage:

split-table [Options] Value

[Options]
-i Feature (Taxa/OTU/Function etc.) table file [Required]
-m Meta data file [Required] (See Meta-data format)
-c Group key [Required]
-o Output file, default is "Out"
-h Help

Example:

split-table –i taxa.txt –m meta.txt –g Status –o taxa.groups

split-matrix

For similarity (distance) matrix by meta-data groups.
The **split-matrix** accepts the results of **comp-sam** and **comp-sam-func**.
Usage:

split-matrix[Options] Value

[Options]
-i Input data matrix [Required]
-m Meta data file [Required] (See Meta-data format)
-c Group key [Required]
-o Output file, default is "Out"
-h Help

Example:

split-matrix –i matrix.txt –m meta.txt –g Status –o matrix.groups

R scripts

PM_Config.R

Install the R package dependency and check environment variable configuration.
Usage:

Rscript PM_Config.R

PM_Distribution.R

For taxa/pathway abundance distribution barchart printing to pdf format. This function has also been integrated in **taxa-sel** and **func-sel** by parameter "-p"
PM_Distribution.R accepst the **relative abundance table** (*.Abd) results of **taxa-sel** and **func-sel**.
Usage:

Rscript PM_Distribution.R [Options] Value

[Options]
-i ABUND_FILE, --abund_file=ABUND_FILE
Input feature table with Relative Abundance [Required]
-o OUTFILE, --outfile=OUTFILE
Output distribution file [default distribution.pdf]
-v THRESHOLD, --threshold=THRESHOLD
Average value threshold [Optional, default 0.01]
-h, --help

Show this help message and exit

PM_Heatmap.R

For heatmap figure printing to pdf format. This function has also been integrated in **comp-sam** and **comp-sam-func** by parameter "-p".

PM_Heatmap.R accepts the results of **comp-sam** and **comp-sam-func**.

Usage:

Rscript PM_Heatmap.R [Options] Value

[Options]

-d DIST_FILE, --dist_file=DIST_FILE
Input distance matrix file [Required].

-o OUTFILE, --outfile=OUTFILE
Output heatmap [default heatmap.pdf]

-h --help
Show this help message and exit

PM_Hcluster.R

For hierarchical clustering and printing to pdf format. This function has also been integrated in **comp-sam** and **comp-sam-func** by parameter "-p".

PM_Hcluster.R accepts the results of **comp-sam** and **comp-sam-func**.

Usage:

Rscript PM_Hcluster.R [Options] Value

[Options]

-d DIST_FILE, --dist_file=DIST_FILE
Input distance matrix file [required].

-o OUTFILE, --outfile=OUTFILE
Output file [default hcluster.pdf]

-k GROUPNUM, --groupNum=GROUPNUM
Number of groups to rect [default 2]

-h --help
Show this help message and exit

PM_Pcoa.R

For PCoA (Principle Co-ordinate Analysis) based on the distance matrix and results printing to pdf format.

PM_PCoa.R accepts the results of **comp-sam** and **comp-sam-func**.

Usage:

Rscript PM_Pcoa.R [Options] Value

[Options]

-d DIST_FILE, --dist_file=DIST_FILE

Input distance matrix file [Required].

-m META_DATA, --meta_data=META_DATA (See Meta-data format)

Input meta-data file [Required].

-l DRAWLABEL, --drawlabel=DRAWLABEL

If enable the sample label [Optional, default FALSE]

-o OUTFILE, --outfile=OUTFILE

Output PCoA [default pcoa.pdf]

-h --help

Show this help message and exit

PM_Pca.R

For PCA (Principle Component Analysis) based on the competent table and results printing to pdf format.

PM_Pca.R accepts the results of **taxa-sel** and **func-sel**.

Usage:

Rscript PM_Pca.R [Options] Value

[Options]

-i ABUND_FILE, --abund_file=ABUND_FILE

Input feature table with Relative Abundance [Required].

-m META_DATA, --meta_data=META_DATA (See Meta-data format)

Input meta-data file [Required].

-l DRAWLABEL, --drawlabel=DRAWLABEL

If enable the sample label [Optional, default FALSE]

-o OUTFILE, --outfile=OUTFILE

Output PCA [default pca.pdf]

-h --help

Show this help message and exit

PM_ADiversity.R

For multivariate statistical analysis of alpha diversity based on the sequence count table.

PM_ADiversity.R accepts the **absolute count table** (*.Count) results of **taxa-sel** and **func-sel**.

Usage:

Rscript PM_Adiversity.R [Options] Value

[Options]

-i COUNT_FILE, --count_file=COUNT_FILE

Input feature table with Absolute Count [Required]

-m META_DATA, --meta_data=META_DATA (See Meta-data format)

Input meta-data file [Required]

-o OUT_DIR, --out_dir=OUT_DIR

Output file name[default Alpha_diversity]

-p PREFIX, --prefix=PREFIX

Output file prefix [Optional, default Out]

-h --help

Show this help message and exit

PM_Bdiversity.R

For multivariate statistical analysis of beta diversity based on the distance matrix.

PM_BDiversity.R accepts the results of **comp-sam** and **comp-sam-func**.

Usage:

Rscript PM_Bdiversity.R [Options] Value

[Options]

-d DIST_FILE, --dist_file=DIST_FILE

Input distance matrix file [Required].

-m MAP_DATA, --map_data=META_DATA (See Meta-data format)

Input meta-data file [Required].

-o OUTDIR, --outdir=OUTDIR

Output directory [default .]

-p PREFIX, --prefix=PREFIX

Output file prefix [Optional, default Out]

-n DIST_NAME, --dist_name=DIST_NAME

The distance metrics name such as Meta-Storms, Jensen-Shannon, Euclidean et al. [Optional, default Default]

-h --help

Show this help message and exit

PM_Marker.R

For bio-marker detection based on ram-sum test.

PM_Marker.R accepts the results of **taxa-sel** and **func-sel**.

Usage:

Rscript PM_Marker.R [Options] Value

[Options]

-i ABUND_FILE, --abund_file=ABUND_FILE

Input feature table with Relative Abundance [Required].

-m META_DATA, --meta_data=META_DATA (See Meta-data format)

Input meta-data file [required].

-o OUTFILE, --outfile=OUTFILE

Output path [default Marker]

-p PREFIX, --prefix=PREFIX

Output file prefix [default Out]

-P PAIRED, --Paired= PAIRED

If paired samples [default FALSE]

-h --help

Show this help message and exit

PM_RFscore.R

For bio-marker scoring using Random Forest algorithm.

PM_RFscore.R accepts the results of **taxa-sel** and **func-sel**.

Usage:

Rscript PM_RFscore.R [Options] Value

[Options]

-i TABLE_FILE, --table_file=TABLE_FILE

Input feature table with read count (*.Count) or abundance (*.Abd) [Required]

-m META_DATA, --meta_data=META_DATA

Input meta-data file [Required]

-o OUT_DIR, --out_dir=OUT_DIR

Output file name[default RFimportance]

-p PREFIX, --prefix=PREFIX

Output file prefix [Optional, default Out]

-h, --help

Show this help message and exit

PM_Network.R

For network based co-occurrence analysis. The output is in pdf format. This function has also been integrated in **comp-corr** by parameter "-N".

PM_Network.R accepts the results of **comp-corr**, **comp-sam** and **comp-sam-func**.

Usage:

Rscript PM_Network.R [Options] Value

[Options]

-i DIST_FILE, --dist_file=DIST_FILE

Input distance table [Required]

-o OUTFILE, --outfile=OUTFILE

Output Network [default network.pdf]

-p POSITIVE_EDGES, --positive_edges=POSITIVE_EDGES

If enable the positive edges [Optional, default TRUE]

-n NEGATIVE_EDGES, --negative_edges=NEGATIVE_EDGES

If enable the negative edges [Optional, default TRUE]

-t THRESHOLD, --threshold=THRESHOLD

Edge threshold [Optional, default 0.7]

-h --help

Show this help message and exit

Input Format

Sequence format and sequence list

Parallel-META accepts sequences in Fasta/Fastq format. For Fasta/Fastq format, sequence labels should not contain space symbol (‘ ‘) and table symbol (‘\t’), and each sequence is in 1 single line. For 16S rRNA sequences, Parallel-META support pair-ended sequences (-R), and for metagenomic shotgun sequences Parallel-META only support single-ended sequences.

For **pipeline,** input path of all input samples (pairs) should be contained in the sequence list. In the sequence list:

Single-ended sequences: each sample has one single line for one Fasta/Fastaq file, such as:

/home/data/sample1.fasta
/home/data/sample2.fasta
/home/data/sample3.fasta

Pair-ended sequences: each sample has two lines for pair-1 sequences and pair-2 sequences in Fasta/Fastq format, such as

/home/data/sample1_pair1.fasta
/home/data/sample1_pair2.fasta
/home/data/sample2_pair1.fasta
/home/data/sample2_pair2.fasta
/home/data/sample3_pair1.fasta
/home/data/sample3_pair2.fasta
/home/data/sample4_pair1.fasta
/home/data/sample4_pair2.fasta

Sample list format

The **class-tax**, **class-func**, **comp-sam**, **comp-sam-func**, **taxa-sel** and **func-sel** accept multiple input samples in sample list format. In the sample list file, each line ONLY contains 1 single sample’s input path.

For class-tax, comp-sam and taxa-sel, the input list contains the taxonomical analysis results’ path (classification.txt, see Results-Single sample);

For class-func, comp-sam-func and func-sel, the input list contains the functional analysis results’ path (functions.txt, see Results-Single sample).

Meta-data format

The **pipeline**, **PM_Pca.R**, **PM_Pcoa.R**, **PM_ADiversity.R**, **PM_BDiversity.R**, **PM_Marker.R** and **PM_RFsocre.R** required the meta-data for all input samples for supervised clustering and bio-marker discovery. Input samples should have the same

order in meta-data file and sample list file. In the meta-data file, each row represents one sample and each column represents one feature. Sample IDs should not be started with number and symbol '#', and should not contain space symbol (' '), backslash symbol ('/') and table symbol ('\t').

Example:

SampleID	Habitat	Sex	Host
Sample1	Palm	M	H1
Sample2	Oral	F	H2
Sample3	Gut	M	H1
Sample4	Gut	F	H3

Visualization

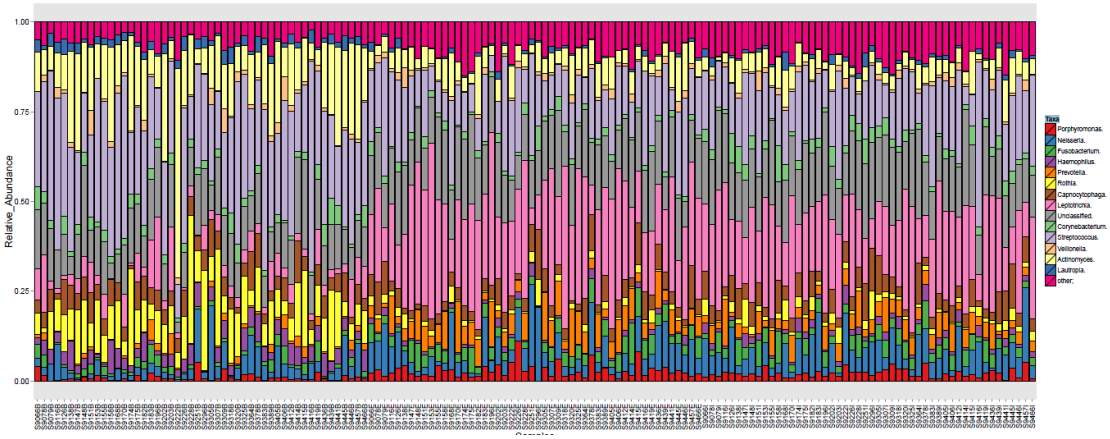
Single sample view

The Single sample view shows the distributions of components on different taxonomy level are displayed in a dynamic pie chart. Click each name at the bottom to search corresponding name from NCBI. This figure is generated by **class-tax**.

Multi-Sample View

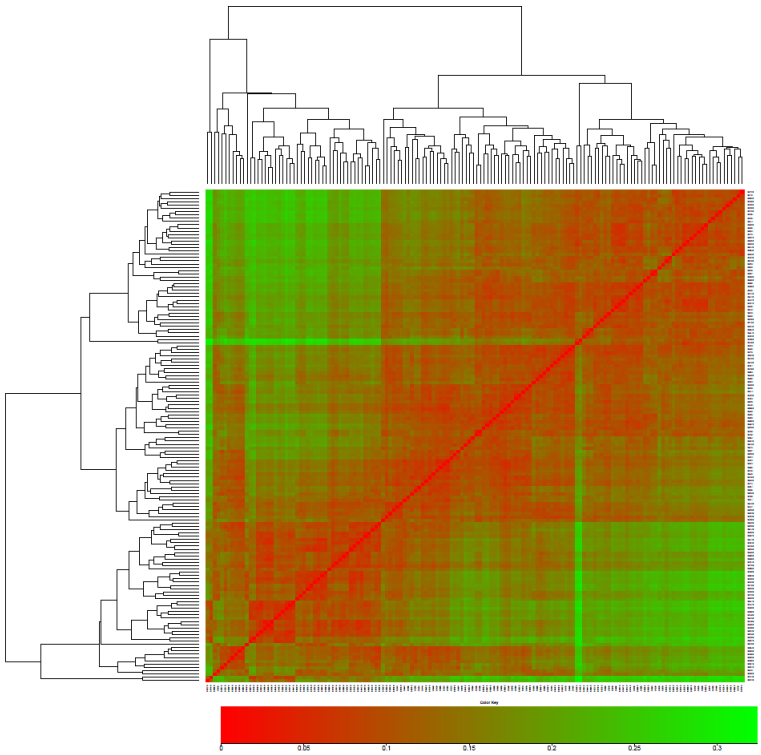
To display the Sample View of all samples into one page.

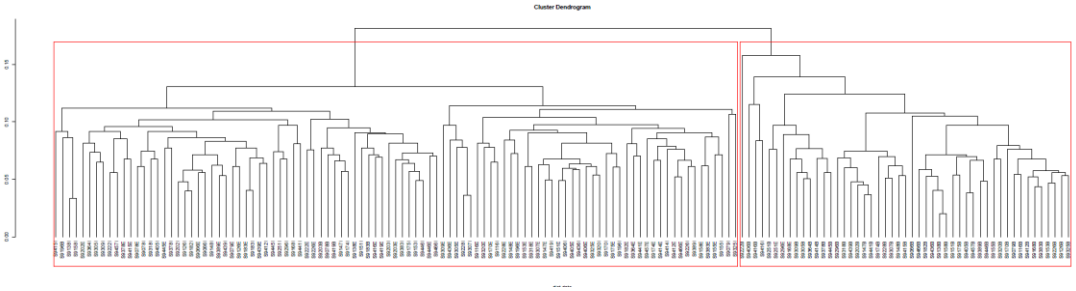
Taxonomy & Function distribution



The distribution bar chart shows the variation of taxa & pathway abundance in each sample. This figure is generated by PM_Distribution.R, and also integrated in **taxa-sel** and **func-sel**.

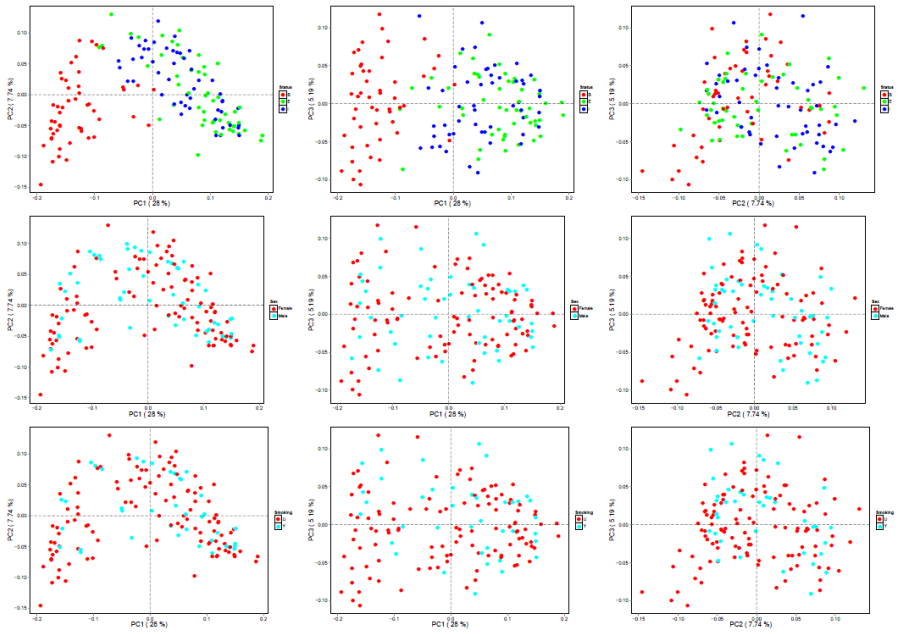
Heatmap & Clustering





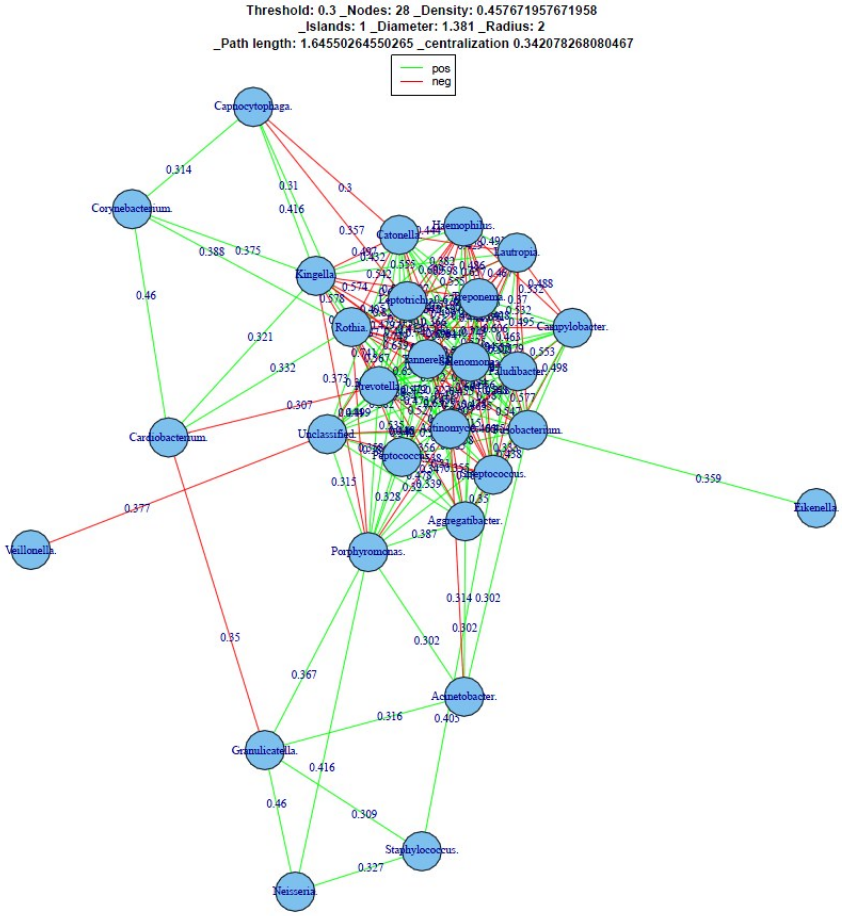
The heatmap shows the quantitative distance and unsupervised hierarchical results of samples. The heatmap figure is generated by **PM_Heatmap.R** and clustering figure is generated by **PM_HClusters.R**. The two functions are also integrated in **comp_sam** and **comp_sam_func**.

PCA & PCoA



The PCA and PCoA results show the supervised clustering results of taxonomical and functional distribution of samples. This figure is generated by **PM_Pca.R** and **PM_Pcoa.R**.

Co-occurrence network



The co-occurrence network shows the co-relationship and the network features among sample taxa / pathway. This figure is generated by **PM_Network.R**, and also integrated in **comp-corr**.

Results

Multiple samples

For multiple sample analysis by Parallel-META pipeline, analysis results will be stored in the results path assigned by parameter –o. In the result folder, there will be the following folders including:

Analysis_Report.txt

The analysis report including parameters configuration and analysis information statistics

Single_Sample

The profiling analysis results of each single sample, which contains the results of Single sample.

Single_Sample.Rare

The rarefied profiling analysis results of each single sample for library size normalization.

Single_Sample.List

The taxonomical and functional result lists of each single sample in Sample list format.

Abundance_Tables

The relative abundance tables (*.Abd) and absolute count tables (*.Count) of multiple samples on different taxonomical and functional levels.

Distance_Matrix

The distance matrix and unsupervised clustering results based on OTUs and KO profiles of multiple samples.

Clustering

The supervised clustering results based on PCA and PCoA with different distance metrics.

Alpha_Diversity

The rarefaction and multivariate statistical analysis results of alpha diversity.

Beta_Diversity

The multivariate statistical analysis results of beta diversity.

Markers

The Biomarker features of different groups on different taxonomical and functional levels.

Network

The microbial interaction network based on different taxonomical and functional levels.

scripts.sh

The detailed scripts of each analysis step.

error.log

The warning and error messages.

Single sample

For single sample analysis of Parallel-META, analysis results will be stored in the result path assigned by parameter –o. In the result folder, there will be:

Analysis_Report.txt

The analysis report including parameters configuration and analysis information statistics

classification.txt

The taxonomy information of the sample in text format.

taxonomy.txt

The taxonomy distribution statistic data of the sample in text format.

functions.txt

The predicted function information of the sample in text format.

meta.rna

The extracted 16S rRNA fragment, if the input is metageomic shotgun sequences.

Notice

1. Please setup the environment variable for Parallel-META before use.
2. For source code package based installation, make sure the dependency packages have been installed to the right directory.
3. The output path should not be the same path as the input file for the output path will be cleared initially. Make sure parallel-meta has the write permission of the output path.
4. The sample ID should not be started by number, and should not contain any “space” or “tab”.
5. We recommend that the maximum length of input sequences should not be longer than 2000bp.
6. Make sure the input is in fasta or fastaq format, and no “space” or “tab” character in sequence id.

Contact

Any problem please contact

Mr. SU Xiaoquan
suxq@qibebt.ac.cn
Mr. JING Gongchao
jinggc@qibebt.ac.cn