Tutorial of Parallel-META 3: How to process your microbiome data

SOP (Standard Operation Procedure) on Parallel-META 3 version 3.5

Version: 3.5
Date: Feb. 20, 2019
Updated by Parallel-META development team

Single-Cell Center
Qingdao Institute of Bioenergy and Bioprocess Technology
Chinese Academy of Sciences

1. How to install Parallel-META

Parallel-MEA 3 is open source, and the executive binary packages for Linux X86-64 and Mac OS X are also provided. For normal users we strongly recommend the executive binary packages with automatics installer to make the installation easier.

Parallel-MEA 3 requires cran R environment to be installed on your computer in advance, which can be freely downloaded from https://www.r-project.org/.

1.1 Download the right package for your operating system

from http://bioinfo.single-cell.cn/parallel-meta.html.
Both source and binary distributions are available.

1.2 Automatic installation

From 3.5 Parallel-META 3 provides an fully automatic installer for easy installation. Here we will install the Parallel-MEA 3 to */opt/tools/*.

```
tar -xzvf parallel-meta.tar.gz   // Unpack the package
mv parallel-meta /opt/tools/   //move the package to your directory
cd /opt/tools/parallel-meta   //change to the parallel-meta directory
source install.sh                    //run the automatic installer
```

Now you have installed Parallel-META 3 in your system! Enter the command

```
PM-pipeline
```

at your terminal and press Enter to check if it works.

Tips for the Automatic installation

1. Please "cd parallel-meta" directory before run the automatic installer.
2. The automatic installer only configures the environment variables to the default configuration files of "~/.bashrc" or "~/.bash_profile". If you want to configure the environment variables to other configuration file please use the manual installation.
3. If the automatic installer failed, Parallel-META 3 can still be installed manually by the following steps in 1.3.

1.3 Manual installation

If the automatic installation failed, you can also install Parallel-META 3 manually. Here we will manually install the Parallel-MEA 3 to */opt/tools/*.

a. Unpack

```
tar -xzvf parallel-meta.tar.gz   // Unpack the package
```

mv parallel-meta /opt/tools/ //move the package to your directory

b. Set the environment variables
The environment variables help Parallel-META to find its databases, and also help you to easily run it in the system. Here we add the following variables in to the default environment variable configuration file "~/.bashrc" (or "~/.bash_profile")by vi.

vi ~/.bashrc

Enter "i" to edit the file, add the following two commands to your "*.bashrc*".

export ParallelMETA=/opt/tools/parallel-meta

export PATH="$PATH:$ParallelMETA/bin"

then press "Esc" and enter ": wq" to save, and active the variables

source ~/.bashrc

c. Install the R packages
Install the depended R packages for statistics and visualization, as follows.

Rscript $ParallelMETA/Rscript/PM_Config.R

d. Compile the source code
If you want to install Parallel-META 3 by source code package, you can build the binaries as follows (for the src package only).

cd /opt/tools/parallel-meta

make

Now you have manually installed Parallel-META in your system! Enter the command

PM-pipeline

at your terminal and press Enter to check if it works.

2. Prepare your samples

Parallel-META accepts both metagenomic shotgun sequences and 16S/18S/ITS rRNA amplicon sequences in FASTA/FASTQ format. In this tutorial we use a case study with 20 16S rRNA amplicon sequencing samples as an example. You can download the example from http://bioinfo.single-cell.cn/parallel-meta.html. In this example the sequence file is "seqs.fa", and meta-data is "meta.txt". Most example commands in this tutorial can be copy-pasted to your terminal for running, such as:

PM-pipeline

Please also notice that some example commands with a * need to be modified on parameter(s) in bold font based on your samples for running, such as:

PM-pipeline ***-i seqs_18s.list*** *-m meta.txt -o out_18s -D E* ***

2.1 Split your sequences.

Sequences must be split into individual samples, which means that each sample is in one single file. If your sequences have already been split into single files, skip this step. PM-split-seq can split all your sequences into single files by their sample barcodes

PM-split-seq ***-i seqs_barcode/seqs_barcode.fa -b seqs_barcode/barcode.txt*** *- o seqs*

or by Mothur format groups file

PM-split-seq ***-i seqs_group/seqs_group.fa -g seqs_group/group.txt*** *-o seqs*

or from QIIME input format sequence file

PM-split-seq ***-i seqs_qiime/seqs_qiime.fa –q T*** *-o seqs*

Then you can find the individual samples in "seqs".

2.2 Make the sequence files list

You should let Parallel-META know the paths of your input samples by a sequences file list. In the list each line contains the exact path of one single sequence file of a sample, so the number of lines in the list should equal the number of your input samples. We strongly recommend the absolute paths (full paths) to avoid the path errors; the relative paths are also supported. In the last step, the *PM-split-seq* automatically saves the list in a file named "seqs.list". We can also manually make the similar input sequence files list. Here is an example of the sequence files list

seqs/sampleA.fa
seqs/sampleB.fa
seqs/sampleC.fa

2.3 Check your meta-data

Meta-data is a table that contains the samples' IDs and information that you want to analysis and compare. In the table, samples should be ordered as in sequence files list of the last step. Each row represents one sample and each column represents one feature. Columns must be separated by Tab symbol ('\t', space symbol ' ' is NOT accepted). The first column is samples' IDs that should NOT be started with number and symbol '#'. All information of sample descriptions in the meta-data table should NOT contain any space symbol (' '), backslash symbol ('/') and table symbol ('\t'). Here is an example of the meta-data table.

ID	Status	Sex	Smoking
SampleA	B	Male	Y
SampleB	B	Male	U
SampleC	B	Female	U

3. A quick start

Now you have prepared all you need to perform the Parallel-META 3. Here we provide a comprehensive, fully automatic and optimized parallel computing pipeline that can fulfill most requirements of microbiome analysis, jut by 1 single line command:

PM-pipeline -i seqs.list -m meta.txt -o out_1

Then what you need to do is just wait for the results, and will not be a long time.

4. Understand the output

After using PM-pipeline, you might get the following folders/files in the output directory. In each directory, files/tables/figures are named with prefix “taxa” are taxonomy results, as well as “func” are metabolic functional results. From 3.4.3 PM-pipeline provides an index page for results browsing.

4.1 index.html (web page)

This is the index page to browse for results browsing. Users can open it by a webpage browser and view the detailed results by hyperlink s (**Figure 1**). Please notice that

a. the “index.html” only works in the output directory;

b. JavaScript, SVG, HTML5 and PDF should be supported by the browser;

c. Links may not available with customized parameters. See "More results" for all available results.

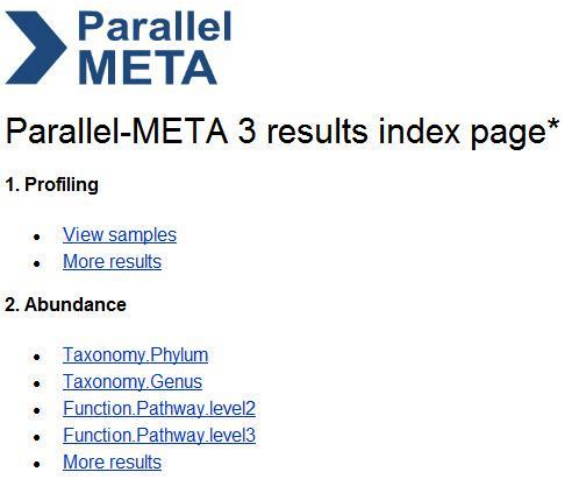

Figure 1. Parallel-META 3 results index page.

4.2 Sample_Views (dir)

This directory contains the visualized sample view (taxonomy.html, JavaScript, SVG and HTML5 should be supported, **Figure 2**) in interactive pie charts across multiple samples.

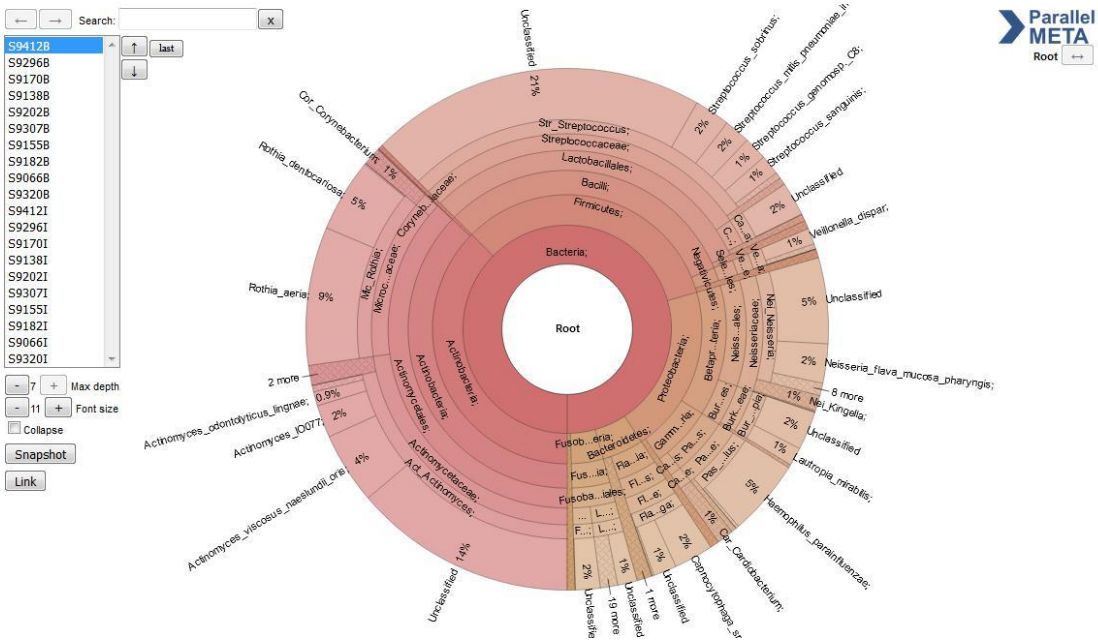

Figure 2. Interactive pie chart of multiple sample view

4.3 Abundance_Tables (dir)

This directory contains the relative abundance tables (*.Abd), absolute sequence count tables (*.Count), and bar charts (*Abd.pdf) of multiple samples on different taxonomical and functional levels (**Figure 3**).

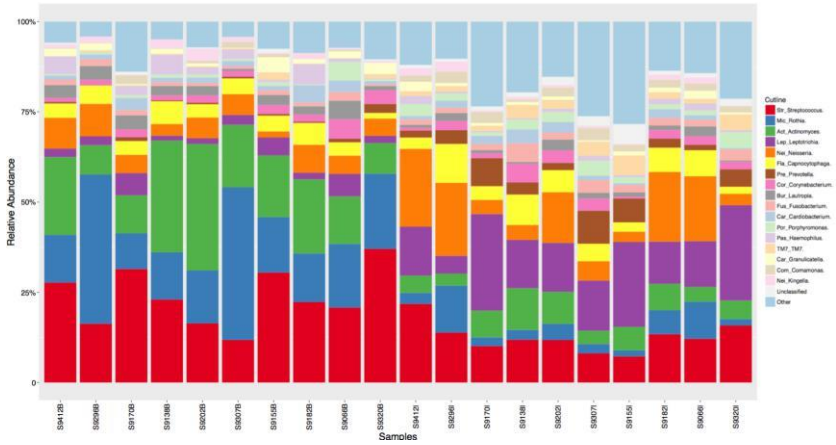

Figure 3. Bar chart on genus level.

4.4 Distance_Matrix (dir)

This directory contains the pair-wised distance matrix (*.dist) of all input samples and unsupervised clustering results (*.dist.clusters.pdf and *.dist.heatmap.pdf) based on OTUs and KO profiles of multiple samples. Distances are computed based on Meta-storms algorithm (Su, et al., *Bioinformatics*, 2012, **Figure 4**).

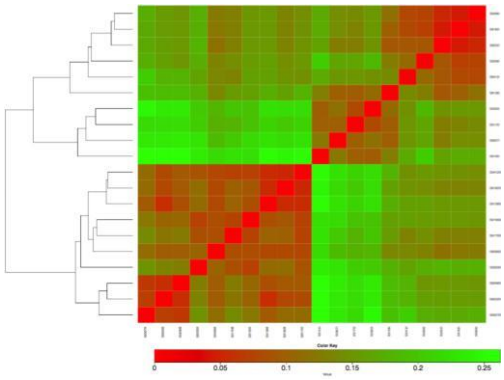

Figure 4. Heatmap and unsupervised clustering of distance matrix

4.5 Clustering (dir)

This directory contains the supervised clustering results based on PCA (*.pca.pdf) and PCoA (*.pcoa.pdf, **Figure 5**) .

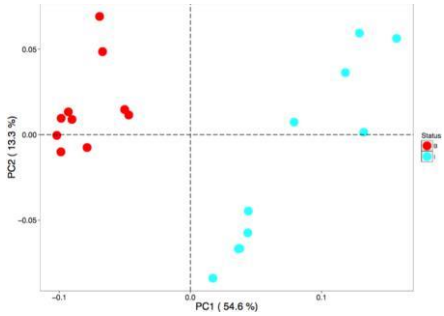

Figure 5. PCoA based supervised clustering

4.6 Alpha_Diversity (dir)

This directory contains the multivariate statistical analysis results (*.Alpha_diversity_Boxplot.pdf and *.Alpha_diversity_Index.txt, **Figure 6**) and rarefaction curve (optional, refer to section 5.6 for details) of alpha diversity. *P*-values are estimated by rank-sum tests.

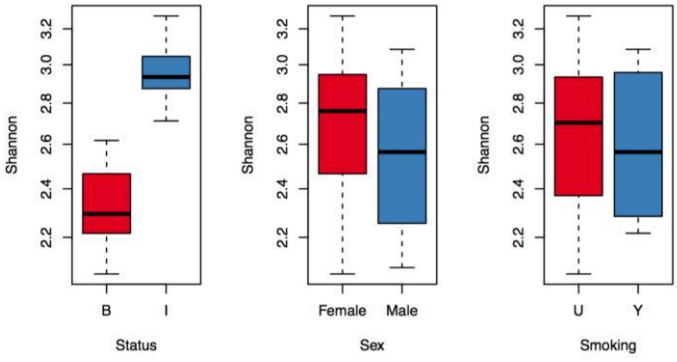

Figure 6. Alpha diversity statistical analysis

4.7 Beta_Diversity (dir)

This directory contains the multivariate statistical analysis results (*.Beta_diversity_Boxplot.pdf and *.taxa.dist.Beta_diversity_Values.txt) of beta diversity. *P*-values are estimated by Adonis/Permanova tests (**Figure 7**).

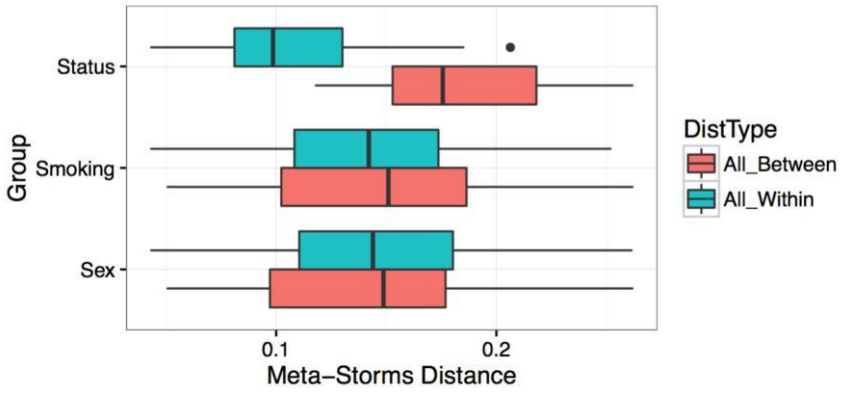

Figure 7. Beta diversity statistical analysis

4.8 Markers (dir)

This directory contains the biomarker organisms (*.sig.boxplot.pdf and .sig.meanTests.xls) and their Random Forest scores (*.RFimportance.pdf and *. RFimportance.txt) among different groups (**Figure 8**).

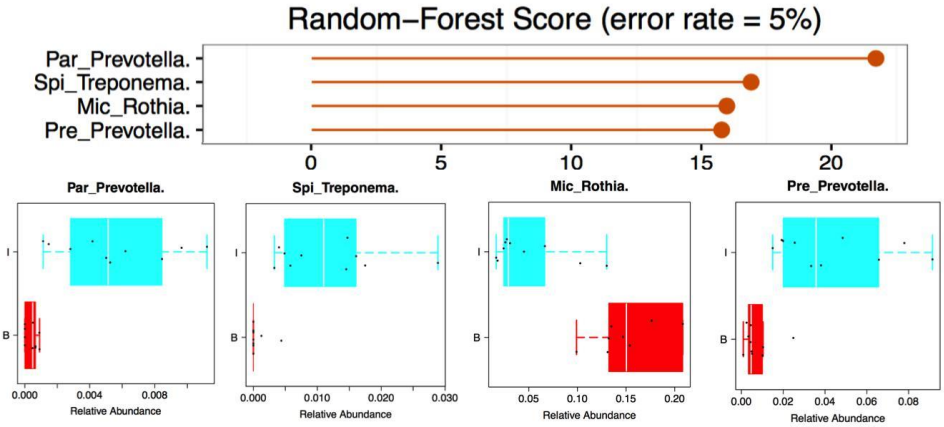

Figure 8. Bio-marker selection

4.9 Network (dir)

This directory contains the microbial interaction network (*.network.pdf) based on different taxonomical and functional levels (**Figure 9**).

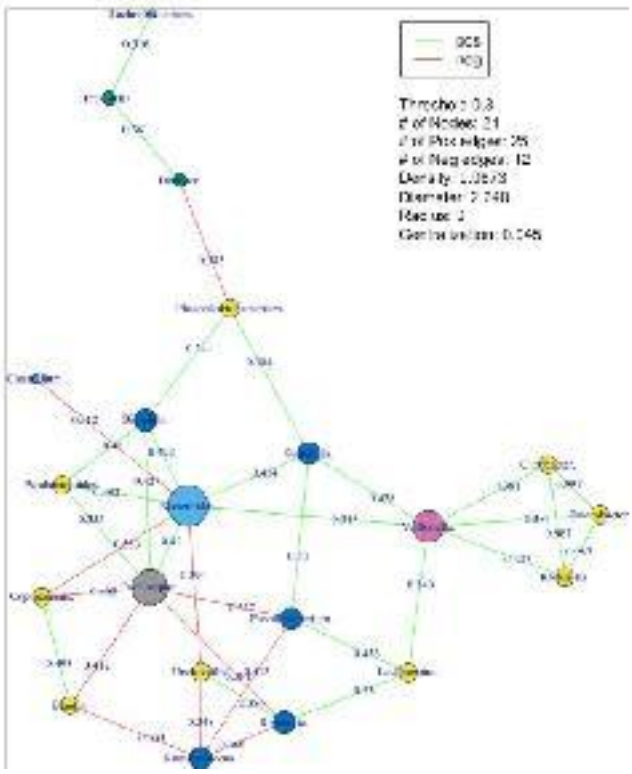

Figure 9. Microbial interaction network

4.10 Single_Sample (dir) and Sinlge_Sample.Rare (dir)

This directory may contain

a. **Single_Sample**: The profiling analysis results of each single sample, which contains the results of Single sample.

b. **Single_Sample.Rare**: The rarefied profiling analysis results of each single sample if the sequencing depth normalization is enabled (-s / -b). Refer to section 5.5 for more details.

In the **Single_Sample/Single_Sample.Rare** directory, each sub directory is the detailed information of an individual sample named by the sample ID. In the sub directories there may be

a. **classification.txt** (plain-text file): The OTUs and taxonomy information of this sample (new version, compatible with 3.4.2 or later).

b. **classification_detail.txt** (plain-text file): The detailed sequence mapping, OTUs and taxonomy information of this sample (compatible with 3.4.1 or lower).

c. **functions.txt** (plain-text file): The predicted KO function information of this sample.

d. **taxonomy.html** (HTML webpage): The visualized sample view in interactive pie chart of this sample.

e. **meta.rna** (fasta sequences): The extracted 16S/18S rRNA fragment, if the input is metageomic shotgun sequences (refer to section 5.3 for details).

f. **Analysis_Report.txt** (plain-text file): The analysis report including parameters configuration and analysis information statistics.

4.11 Single_Sample.List (dir)

This directory contains the taxonomical and functional results path lists (*.list) of all samples. Each list has 2 columns: the first column is the samples' ID and the second column is the path of the taxonomical/functional results. In this directory normally there are 2 files:

a. **taxa.list**: the taxonomical results list

b. **func.list:** the functional results list

and if the sequencing depth normalization is enabled (-s / -b, refer to section 5.5) you can also find:

c. **taxa.rare.list:** the rarefied taxonomical results list

4.12 Logs (plain-text file)

a. **Analysis_Report.txt**: The analysis report including parameters configuration and analysis information statistics.

b. **scripts.sh**: The detailed scripts of each analysis step.

c. **error.log**: The warning and error messages.

5. Advanced parameters

You can also modify and customize your analysis process. Here are some commonly used parameters:

5.1 16S/18S/ITS rRNA sequences

Parallel-META accepts 16S/18S/ITS rRNA sequences. The default reference database is GreenGenes-13-8 and you can use parameter "-D" to change reference database ("-D G" use GreenGenes-13-8 (16S rRNA, 97% level), "-D E" use SILVA (18S rRNA, 97% level), "-D O" use Oral_Core (16S rRNA, 97% level) and "-D T" use ITS (ITS1, 97% level)).

*PM-pipeline -i seqs_18s.list -m meta.txt -o out_18s -D E/O/T **

In addition, please make all samples in the same list have the same sequence type (all are 16S rRNA or all are 18S rRNA or all are ITS rRNA)

5.2 Pair-ended 16S/18S rRNA sequences

Parallel-META accepts 16S/18S rRNA sequences in pair-end format. To use pair-ended sequences, make each pair in a separated FASTA/FASTQ file, and then put 2 file paths for each sample in the sequence list (seqs.list). For example, the sinlge-ended input sequence list (contains 3 samples) likes

seqs/sampleA.fa
seqs/sampleB.fa
seqs/sampleC.fa

and the pair-ended sequence list (also contains 3 samples, and each sample in 2 lines) likes

seqs/sampleA_end1.fa
seqs/sampleA_end2.fa

seqs/sampleB_end1.fa
seqs/sampleB_end2.fa
seqs/sampleC_end1.fa
seqs/sampleC_end2.fa

In addition, a parameter "-P" is available to indicate the orientation of the pair-ended sequences:

-P 0: Forward & Reverse (default)
-P 1: Forward & Forward
-P 2: Reverse & Forward

PM-pipeline ***-i seqs_pair.list*** *-m meta.txt -o out_pair* ***-P 0 ****

5.3 Metagenomic Shotgun sequences

Parallel-META accepts metagenomic shotgun sequences as input by extracting the 16S/18S rRNA fragments contained in the shotgun sequences for profiling. Currently only single-ended shotgun sequences are supported. The shotgun sequence files should in the same format as that of 16S/18S rRNA sequences, and make the directories of your input sequence files in list (seqs.list), then add the parameter "-M T" to indicate the input as shotgun sequences.

PM-pipeline -i seqs.list -m meta.txt -o out_1 ***-M T***

5.4 Re-analysis without profiling

The profiling always takes a long running time, so if all samples have already been profiled into OTUs ("classification.txt"), only the sample list (Single_Sample.List/taxa.list) and meta-data (meta.txt) are needed to preform the re-analysis with adjustable parameters in the following sections

PM-pipeline ***-l out_1/Single_Sample.List/taxa.list*** *-m meta.txt -o out_2*

5.5 Add/remove/change samples

You can easily add, remove or change samples by modify the sequence list / sample list and the corresponding meta-data. For example

a. To add a sample sequence file in the analysis, you need to add the sequence file path in sequence list (seqs.list), and add its meta-data information in meta.txt, then run the

PM-pipeline ***-i seqs.list -m meta.txt*** *-o out_modify*

b. To add a sample OTUs file in the analysis, you need to add the "classification.txt" file path in the sample list (Single_Sample.List/taxa.list) and meta-data information in meta.txt, then run the

PM-pipeline ***-l out_1/Single_Sample.List/taxa.list*** *-m meta.txt -o out_modify*

5.6 Sequencing depth normalization

You need to normalize the sequencing depth of all samples to avoid the bias in diversity analysis if the sequence number varies largely among samples (eg. varies >

5 times). You can either add the normalization in PM-pipeline automatically by parameter -s (sequence number) and -b (bootstrap, optional):

PM-pipeline -i seqs.list -m meta.txt -o out_1 ***-s 1000 -b 200***

PM-pipeline -l out_1/Single_Sample.List/taxa.list -m meta.txt -o out_2 ***-s 1000 -b 200***

or manually run

PM-rand-rare -l out_1/Single_Sample.List/taxa.list -o Single_Sample.Rare ***-s 1000 -b 200***

5.7 Rarefaction curve

The rarefaction curve shows how the alpha diversity varies with the sequence number. You can either add parameter -R T in the PM-pipeline to enable the rarefaction curve

PM-pipeline -i seqs.list -m meta.txt -o out_1 ***-R T*** *PM-pipeline -l*

out_1/Single_Sample.List/taxa.list -m meta.txt -o out_2 ***-R T***

or manual run

PM-rare-curv -i out_1/Abundance_Tables/taxa.OTU.count -o Rare_curve

5.8 More taxonomy levels

By default configuration the Parallel-META only parse out the taxonomy profiles on Phylum and Genus level, and you can add more levels by option -L. Here are the taxonomy level parameters for -L:

Taxonomy level	**Parameter**
Phylum	1 or P
Class	2 or C
Order	3 or O
Family	4 or F
Genus	5 or G
Species	6 or S

For example, you need Phylum, Class, Family and Genus,

PM-pipeline -i seqs.list -m meta.txt -o out_1 ***-L 1245*** *PM-pipeline -l*

out_1/Single_Sample.List/taxa.list -m meta.txt -o out_2 ***-L PCFG***

5.9 Path prefix for list

We strongly recommend the absolute path (full path) in the lists (sequence list: seqs.list and sample lists: Single_Sample.List/taxa.list), which can avoid the path

error when changing the current work directory. However, when you use relative path, and you can add a path prefix by parameter "-p" to make the relative path in the list to be always accessible. This will add the prefix to all paths in the list file.

PM-pipeline -i seqs.list -m meta.txt -o out_1 ***-p /home/data/*** ***

Then Parallel-META will add the prefix "/home/data/" to all paths in seqs.list.

5.10 Thread number for parallel computing

Parallel-META always automatically and dynamically distributes computing tasks to all CPU cores. You can also use parameter "-t" to manually assign the thread number.

PM-pipeline -i seqs.list -m meta.txt -o out_1 -t 8

6. What Parallel-META have done: detailed steps

If you understand all above, then we can go through the details of the pipeline step by step showed in **Figure 10** with more adjustable parameters for customization.

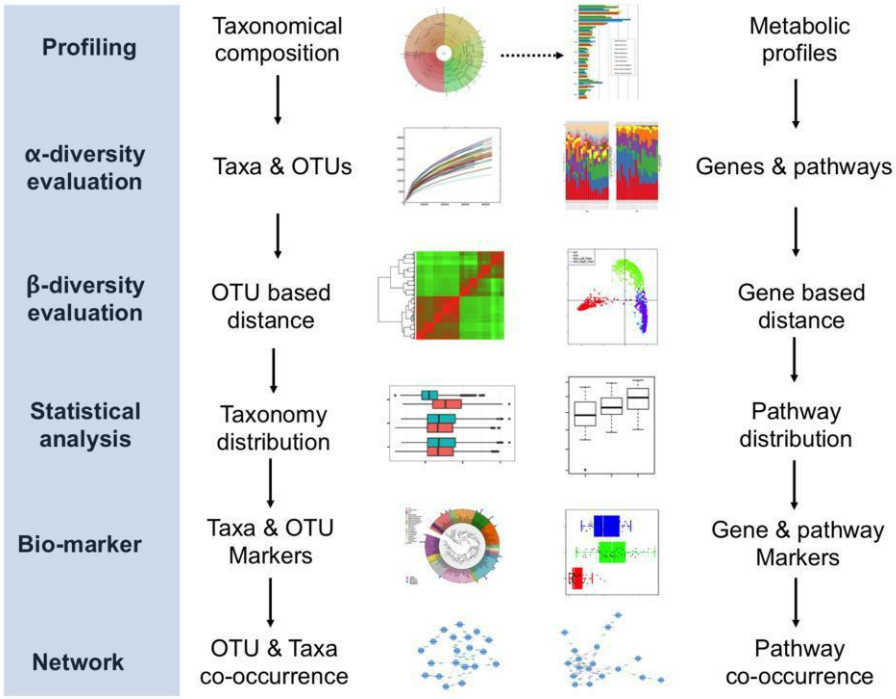

Figure 10. Parallel-META workflow

6.1 Profile a single sample for taxonomy and function

Parallel-META allows you to process a single sample by the FASTA/FASTQ sequence file. The sequences can be either 16S/18S rRNA sequences or metagenomic shotgun sequences.

PM-parallel-meta -r seqs/S9066B.fa -o S9066B.out //16S rRNA

PM-parallel-meta -r seqs/S9066B.fa ***-R seq/S9066B_pair.fa*** *-o S9066B.out* ***
//Pair-ended 16S rRNA

PM-parallel-meta ***-r seqs/S9066B.fa*** *-o S9066B.out* ***-D E **** *//18S rRNA PM-parallel-meta* ***-m seqs/S9066B.fa*** *-o S9066B.out //Shotgun (Bacteria)*

PM-parallel-meta ***-m seqs/S9066B.fa*** *-o S9066B.out* ***-D E **** *//Shotgun (Eukaryote)*

You can find the output files introduced as section 4.9. More available parameters please check the users' manual or use

PM-parallel-meta -h

6.2 Parse out the taxa/function abundance tables

You can parse out abundance tables on any taxonomy/function level by the sample list (Sample_List/taxa.list or Sample_List/func.list) or the OTU/KO count table(taxa.OTU.Count or func.KO.Count). The taxonomy level parameter -L is the similar usage as section 5.7, but here you can only input one single level for one run.

PM-taxa-sel -l out_1/Single_Sample.List/taxa.list -L G -o taxa.genus //taxonomy by list

PM-taxa-sel -T out_1/Abundance_Tables/taxa.OTU.Count -L G -o taxa.genus //taxonomy by OTU table

PM-func-sel -l out_1/Single_Sample.List/func.list -L 2 -o func.l2 //function by list

PM-func-sel -l out_1/ Abundance_Tables/taxa.KO.Count -L 2 -o func.l2 //function by KO table

Then you will get *.Count (sequence count) and *.Abd (relative abundance) as output.

The *PM-taxa-sel* provides 5 parameters for filtering:

-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%)

For OTU level (-L 7) we suggest "-m 0 -n 0 -z 0 -v 0"; for other taxonomy levels (-L 1/2/3/4/5/6) we suggest "-m 0 -n 0". You can adjust these parameters based on your experiment design and data quality. The *PM-func-sel* does not support filtering parameters.

You can add "-P T" for *PM-taxa-sel* and *PM-func-sel* to print out the bar chart on the selected level, or manually run

Rscript $ParallelMETA/Rscript/PM_Distribution.R -i taxa.genus.Abd -o taxa.genus.Abd.pdf

More available parameters please check the users' manual or

use *PM-taxa-sel -h*

PM-func-sel -h

Rscript $ParallelMETA/Rscript/PM_Distribution.R -h

6.3 Alpha diversity: examine the complexity of samples

Parallel-META uses abundance table (*.Abd) to evaluate the alpha diversity of multiple samples among different groups.

Rscript $ParallelMETA/Rscript/PM_Adiversity.R -i out_1/Abundance_Tables/taxa.genus.Abd -m meta.txt -o alpha_diversity_out

More available parameters please check the users' manual or use

Rscript $ParallelMETA/Rscript/PM_Adiversity.R -h

6.4 Beta diversity: compare multiple samples

Parallel-META examines the beta diversity based on the distance matrix (*.dist) calculated by the OTU table or KO table. To generate the pair-wised distance matrix using OTU able (only OTU level is accepted for taxonomy):

PM-comp-sam -T out_1/Abundance_Tables/taxa.OTU.Count -o taxa.dist -d T and using the functional KO table (only KO level is accepted for function):

PM-comp-sam-func -T out_1/Abundance_Tables/func.KO.Count -o func.dist -d T

Then you can use the output taxa.dist for beta diversity analysis among different groups:

Rscript $ParallelMETA/Rscript/PM_Bdiversity.R -d taxa.dist -m meta.txt -o beta_diversity_out

You can also compare any two samples by their OTU/taxonomy information (classification.txt) or function information (functions.txt)

PM-comp-sam -i out_1/Single_Sample/S9066B/ classification.txt out_1/Single_Sample/S9066I/classification.txt -d T

PM-comp-sam-func -i out_1/Single_Sample/S9066B/ functions.txt out_1/Single_Sample/S9066I/functions.txt -d T

More available parameters please check the users' manual or

use *PM-comp-sam -h*

PM-comp-sam-func -h

Rscript $ParallelMETA/Rscript/PM_Bdiversity.R -h

6.5 Clustering: unsupervised and supervised

Parallel-META supports unsupervised clustering (hierarchical clustering) and supervised clustering (PCA and PCoA).

Hierarchical clustering uses distance matrix (*.dist) as input, and output the clustering results and heat map.

Rscript $ParallelMETA/Rscript/PM_Hcluster.R -d

out_1/Distance_Matrix/taxa.dist -o taxa.dist.clusters.pdf

Rscript $ParallelMETA/Rscript/PM_Heatmap.R -d

out_1/Distance_Matrix/taxa.dist -o taxa.dist.heatmap.pdf

For supervised clustering, the PCA takes the abundance tables (*.Abd) as input, and the PCoA takes distance matrix (*.dist) as input.

Rscript $ParallelMETA/Rscript/PM_Pca.R -i

out_1/Abundance_Tables/taxa.genus.Abd -m meta.txt -o taxa.genus.pca.pdf

Rscript $ParallelMETA/Rscript/PM_Pcoa.R -d

out_1/Distance_Matrix/taxa.dist -m meta.txt -o taxa.dist.pcoa.pdf

More available parameters please check the users' manual or use

Rscript $ParallelMETA/Rscript/PM_Hcluster.R -h

Rscript $ParallelMETA/Rscript/PM_Heatmap.R -

h Rscript $ParallelMETA/Rscript/PM_Pca.R -h

Rscript $ParallelMETA/Rscript/PM_Pcoa.R -h

6.6 Bio-marker analysis: determine the key organisms/genes among multiple samples

For bio-marker analysis, Parallel-META firstly selects all organisms that un-evenly distributed among different groups based on the abundance table (*.Abd):

Rscript $ParallelMETA/Rscript/PM_Marker.R -i

out_1/Abundance_Tables/taxa.genus.Abd -m meta.txt -o marker_out

Then all candidate markers should be ranked by Random Forest scores:

Rscript $ParallelMETA/Rscript/PM_RFscore.R -i

out_1/Abundance_Tables/taxa.genus.Abd -m meta.txt -o marker_out

More available parameters please check the users' manual or use

Rscript $ParallelMETA/Rscript/PM_Marker.R -h

Rscript $ParallelMETA/Rscript/PM_RFscore.R -h

6.7 Network analysis: Construct the co-occurrence & co-exclusion network among multiple samples

Parallel-META constructs the microbial interactive network based on the correlation among organisms calculated by the abundance table (*.Abd). Firstly you need to calculate the correlation matrix:

PM-comp-corr -i out_1/Abundance_Tables/taxa.genus.Abd -o taxa.genus.corr

You can add "-N T" for PM-comp-corr to automatically generate out the network, or manually construct the network.

Rscript $ParallelMETA/Rscript/PM_Network.R -i

taxa.genus.corr.self_matrix.out -o taxa.genus.network.pdf

7. Contact

Any problem please contact Parallel-META development team

Mr. JING Gongchao E-mail: jinggc@qibebt.ac.cn
Miss. Zhang Yufeng E-mail: qdu_zyf@163.com