MTD: Meta-Transcriptome Detector

MTD is a software that has two sub-pipelines to detect and quantify microbiomes by analyzing bulk RNA-seq data and single-cell RNA-seq data, respectively. It supports comprehensive microbiome species and vectors, including viruses, bacteria, protozoa, fungi, plasmids, and vectors. MTD is executed in Bash in GNU/Linux system. Users can easily install and run MTD using only one command and without requiring root privileges. The outputs (graphs, tables, count matrixes, etc.) are automatically generated and stored in the designated directory/folder defined by the user.

Requirements

- 160 Gb RAM
- 560 Gb storage space
- Conda was installed
- GNU/Linux system with Bash

Installation

1. Download directly or git clone MTD (git clone https://github.com/FEI38750/MTD.git) to the place you want to install. The full version of software (~530Gb) will be installed in this MTD folder.
2. In termial, type

bash [path/to/MTD]/Install.sh -t [threads] -p [path/to/conda]
For example:

```
bash ~/MTD/Install.sh -t 20 -p ~/miniconda3
```

Notes

- Installation may take 1-2 days.
- If conda hasn't been installed in your system, please use the code below to install the conda:

  ```
  URL=https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
  curl $URL > miniconda-installer.sh
  bash miniconda-installer.sh -b
  ```

  Conda then will be installed in your home directory, such as path: ~/miniconda3
- MTD conda environments occupies ~24Gb in your conda folder.
- Tips: file management software such as FileZilla (https://filezilla-project.org/download.php?show_all=1) can help you to manage your files on HPC/server.

Run MTD

Bulk RNA-seq

1. Store all the paired-end fastq files (accepted: fastq, fastq.gz, fq, fq.gz) to be analyzed in a folder, subfolders for each sample are accepted.
   The paired fastq files must be named starting with the sample name followed by "_1" and "_2". For example, sample1_1.fq.gz and sample1_2.fq.gz are paired-end fastq files for sample1.
2. Prepare the samplesheet.csv. You can copy and modify the one in MTD folder.

Store with name: **samplesheet.csv**

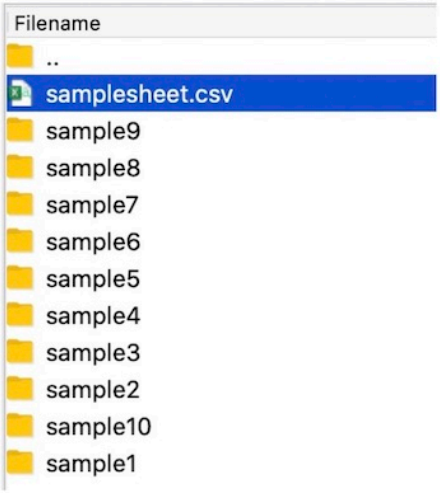
Don't change headers

Modify the body of the sheet

sample_name	group		comparisons	group1	vs	group2
sample1	Treatment1			Treatment1	vs	Control
sample2	Treatment1			Treatment2	vs	Control
sample3	Treatment1					
sample4	Treatment2					
sample5	Treatment2					
sample6	Treatment2					
sample7	Treatment2					
sample8	Control					
sample9	Control					
sample10	Control					

Enter the group comparisons you would like to analyze here

3. Put samplesheet.csv in the same folder as the fastq files.



Put samplesheet.csv in the same folder as the fastq files

4. In termial, type

bash [path/to/MTD]/MTD.sh -i [path/to/samplesheet.csv] -o [path/to/output_folder] -h [host species taxonomy ID] -t [threads]

Host species taxonomy ID: human:9606, mouse:10090, rhesus monkey:9544

For example:

```
bash ~/MTD/MTD.sh -i ~/raw_data/samplesheet.csv -o ~/MTD_output -h 9544 -t 20
```

Single-cell RNA-seq

1. Put the count matrix of host genes in a folder named with the sample name. In this folder, 10x should be a matrix.mtx, a genes.tsv, and a barcodes.tsv; or a single .h5 file. Dropseq should be a .dge.txt file.
2. Type this folder path into the column host_matrix_folder of the samplesheet_SC.csv. For example:

List the path to fastq files ↓

Path to the folder contain the Count Matrix for each sample in 1st column. Folder name is the sample name ↓

fastq_files	host_matrix_folder
~/SC_Raw/SRR4210_R1.fastq	~/SC_folder/sc_OBrain1
~/SC_Raw/SRR4210_R2.fastq	~/SC_folder/sc_OBrain1
~/SC_Raw/SRR4211_R1.fastq	~/SC_folder/sc_OBrain3
~/SC_Raw/SRR4211_R2.fastq	~/SC_folder/sc_OBrain3
~/SC_Raw/SRR4212_R1.fastq	~/SC_folder/sc_YBrain1
~/SC_Raw/SRR4212_R2.fastq	~/SC_folder/sc_YBrain1
~/SC_Raw/SRR4213_R1.fastq	~/SC_folder/sc_YBrain3
~/SC_Raw/SRR4213_R2.fastq	~/SC_folder/sc_YBrain3

e.g., sample name sc_OBrain1 for SRR4210_R1.fastq and SRR4210_R2.fastq

3. In termial, type

bash [path/to/MTD]/MTD_singleCell.sh -i [path/to/samplesheet_SC.csv] -o [path/to/Output_folder] -h [Host species taxonomy ID] -t [Threads] -p [Platform] -d [prime Direction]

Single cell RNAseq platform(-p): enter 1 for 10x or 2 for Dropseq platform

prime_direction(-d): specifying barcode locations: enter 3 or 5 for barcodes are at the 3’ end or 5' end of the read

For example:

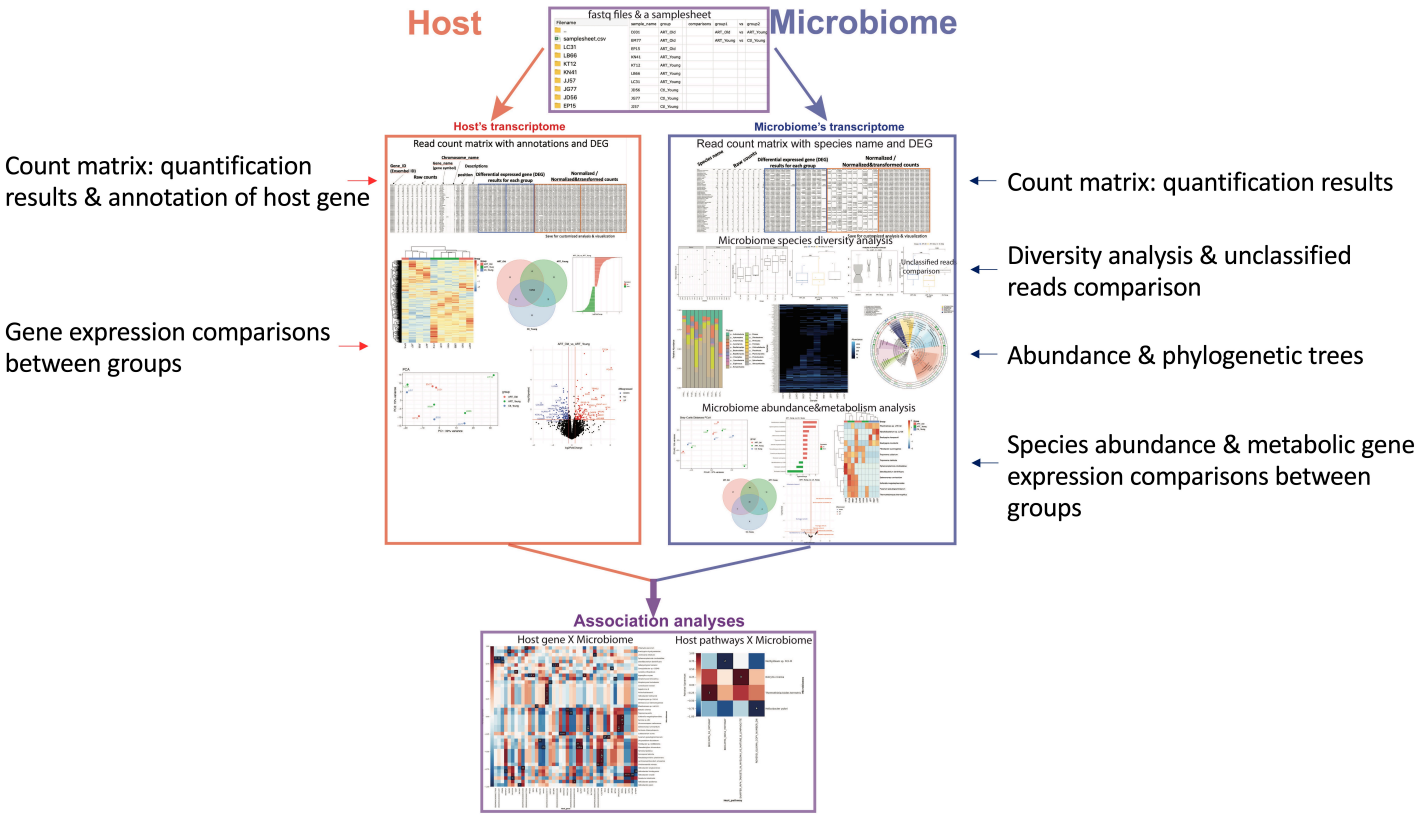
```
bash ~/MTD/MTD_singleCell.sh -i ~/scRNAseq_rawData/samplesheet_SC.csv -o ~/output -h 10090 -t 20 -p 1 -d 3
```

Notes

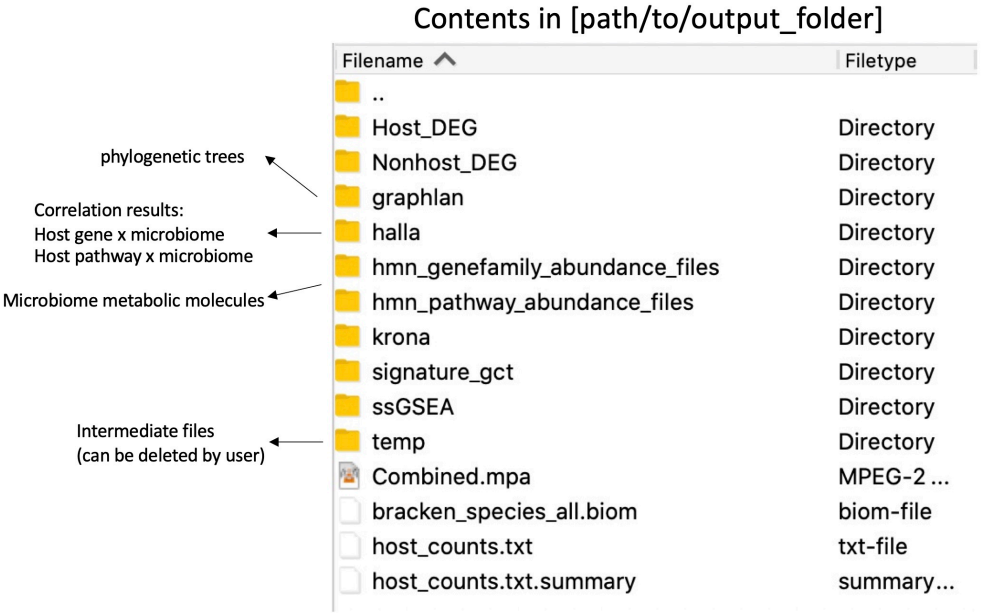
- 10x and Dropseq use paired end sequence. The first fastq file contains barcodes (e.g., 26bp length in SRR4210_R1.fastq). The second fastq file contains transcript's sequences (e.g., 98bp length in SRR4210_R2.fastq).
- Default QC is *subset= nFeature_RNA>200 & nFeature_RNA < 2*median(number_of_Feature_RNA) & percent.mt < 10*
  In addition, user can customize QC by adding -l [Minimum nFeature_RNA] -r [Maximum nFeature_RNA] -m [percent.mt]

Outputs

Bulk RNA-seq



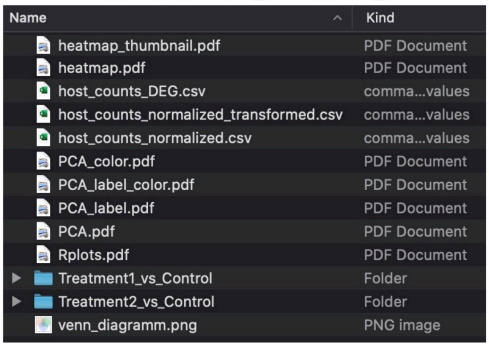
The results are generated automatically and saved in the output folder defined by the user.
The output included:



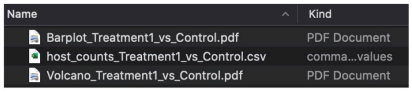
- For **host**: [path/to/output_folder]/Host_DEG/
  The count matrix (host_counts_DEG.csv) contains the Ensembl gene ID, gene symbol, chromosome name, gene position, functional descriptions, DEG results for each group comparison, raw read counts, normalized reads count, normalized and transformed reads counts. This comprehensive count matrix

facilitates the user to perform downstream analyses such as pathway enrichment and customized data visualization.
The data visualization includes the heatmap (with/without gene name), Venn Diagram, PCA, barplot, and volcano plots.
The individual group comparison results are saved in the corresponding subfolder (e.g., group1_vs_group2).

Contents in [path/to/output_folder]/Host_DEG folder

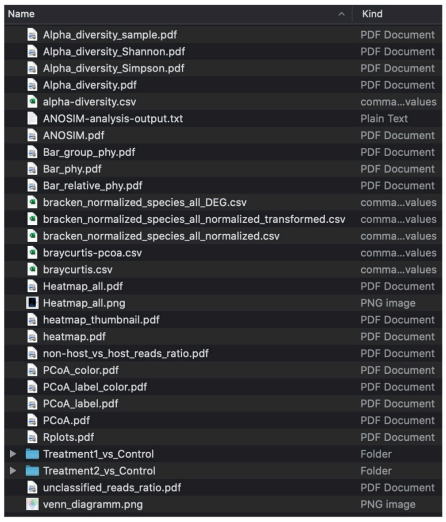


Contents in Treatment1_vs_Control folder

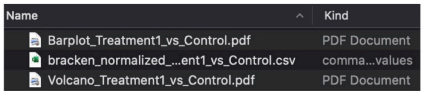


- For **microbiome**: [path/to/output_folder]/Nonhost_DEG/
  The count matrix (bracken_normalized_species_all_DEG.csv) contains the name and taxonomy ID of microbiome species, DEG results for each group comparison, raw read counts, normalized reads count, normalized and transformed reads counts.
  Diversity analysis, unclassified reads comparison, abundance&DEG heatmaps, phylogenetic trees.
  Venn Diagram, heatmap, PCoA, barplot, and volcano plots for the results of species abundance and group comparisons.

Contents in [path/to/output_folder]/Nonhost_DEG folder



Contents in Treatment1_vs_Control folder



- **Microbiome metabolic molecules**: hmn_genefamily_abundance_files contain microbiome metabolic molecules and group comparison results. Results are translated to kegg and go terms to facilitate reading and demonstrated via Venn Diagram, heatmap, PCA, barplot, and volcano plots and count matrix. hmn_pathway_abundance_files contain pathway results of those molecules.
- **Association analysis**: halla folder contains the results of association between:
  host gene and microbiome species
  host pathways and microbiome species

Single-cell RNA-seq

Count matrix for the single-cell microbiome is automatically generated and saved in the output folder.

Microbiome name & taxid	Quantification in single cells							
Name	AAACCTGGTCCTCCAT	AAACGGGAGCTCCCAG	AAACGGGTCAACGCTA	AAAGCAACAAGGACAC	AAAGCAACAAGTAGTA	AAAGCAACAAGTTGTC	AAAGCAAGTCACACGC	AAAGCAAGTGTAATGA
[Candida] glabrata (taxid 5478)	0	0	0	0	0	0	0	0
[Clostridium] innocuum (taxid 1522)	0	0	0	0	0	0	0	0
[Eubacterium] cellulosolvens 6 (taxid 633697)	0	0	0	0	0	0	0	0
[Ruminococcus] gnavus ATCC 29149 (taxid 411470)	0	0	0	0	0	0	0	0
Abelson murine leukemia virus (taxid 11788)	0	0	0	0	0	0	0	1
Acanthamoeba polyphaga moumouvirus (taxid 1269028)	0	0	0	0	0	0	0	0
Acetobacter (taxid 434)	0	0	0	0	0	0	0	0
Acetobacter aceti (taxid 435)	0	0	0	0	0	0	0	0
Acetobacter orientalis (taxid 146474)	0	0	0	0	0	0	0	0
Acetobacter oryzifermentans (taxid 1633874)	0	0	0	0	0	0	0	0
Acetobacter pasteurianus (taxid 438)	0	0	0	0	0	0	0	0
Acetobacter tropicalis (taxid 104102)	0	0	0	0	0	0	0	0
Acetobacteraceae (taxid 433)	0	0	0	0	0	0	0	0
Achromobacter (taxid 222)	0	0	0	0	0	0	0	0
Achromobacter denitrificans (taxid 32002)	0	0	0	0	0	0	0	0
Achromobacter insolitus (taxid 217204)	0	0	0	0	0	0	0	0
Achromobacter sp. B7 (taxid 2282475)	0	0	0	0	0	0	0	0
Achromobacter sp. MFA1 R4 (taxid 1881016)	0	0	0	0	0	0	0	0
Achromobacter spanius (taxid 217203)	0	0	0	0	0	0	0	0
Achromobacter xylosoxidans (taxid 85698)	0	0	0	0	0	0	0	0
Achromobacter xylosoxidans A8 (taxid 762376)	0	0	0	0	0	0	0	0
Acidianus ambivalens (taxid 2283)	0	0	0	0	0	0	0	0
Acidianus brierleyi (taxid 41673)	0	0	0	0	0	0	0	0
Acidianus manzaensis (taxid 282676)	0	0	0	0	0	0	0	0
Acidilobus sp. 7A (taxid 1577685)	1	0	0	0	0	0	0	0
Acidiphilium (taxid 522)	0	0	0	0	0	0	0	0
Acidipropionibacterium acidipropionici (taxid 1748)	0	0	0	0	0	0	0	0
Acidithiobacillus ferrivorans SS3 (taxid 743299)	0	0	0	0	0	0	0	0
Acidobacterium capsulatum ATCC 51196 (taxid 240015)	0	0	0	0	0	0	0	0
Aciduliprofundum sp. MAR08-339 (taxid 673860)	0	0	0	0	0	0	0	0

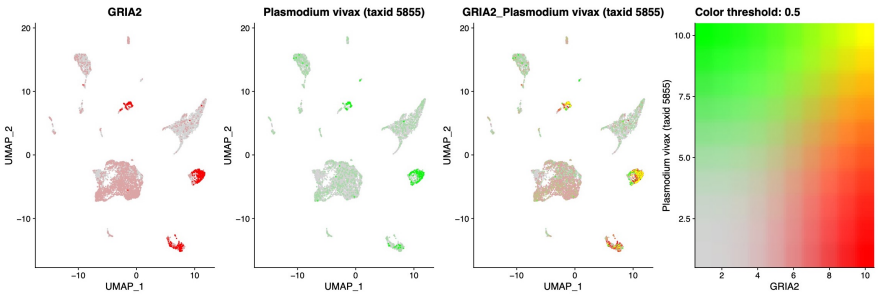
The results of correlation test between microbiome and host genes are generated automatically and saved in the output folder defined by the user.

Contents in [path/to/output_folder]

Filename	Filetype
..	
sc_OBrain3_count_matrix.txt	txt-file
sc_YBrain3_count_matrix.txt	txt-file
sc_YBrain1_count_matrix.txt	txt-file
sc_OBrain1_count_matrix.txt	txt-file
Correlations_MicrobiomeXhost	Directory

Contents in [path/to/output_folder/Correlations_MicrobiomeXhost]

Filename	Filetype
..	
10_highest_cor.pdf	pdf-file
9_highest_cor.pdf	pdf-file
8_highest_cor.pdf	pdf-file
7_highest_cor.pdf	pdf-file
6_highest_cor.pdf	pdf-file
5_highest_cor.pdf	pdf-file
4_highest_cor.pdf	pdf-file
3_highest_cor.pdf	pdf-file
2_highest_cor.pdf	pdf-file
1_highest_cor.pdf	pdf-file
MicrobiomeXhost_sigCorrelation.tsv	tsv-file



Notes

- For reference, the MTD running time would be:
  - Bulk RNA-seq: for 10 samples, 20 fastq files (total 47 Gb) by using 20 threads CPU is ~8 hours (except correlation analysis). In addition, the correlation analysis may need a further ~26-30hours (results in halla folder). So the total running time would be ~34-38 hours.
  - Single-cell RNA-seq: for fastq files contain ~2000 cells, by using 20 threads CPU is ~2 hours.
- Users can add any other host species easily by one command line:

bash Customized_host.sh -t [threads] -d [host_genome_Ensembl_address] -c [host_taxid] -g [host_gtf_Ensembl_address]

For example:

`bash ~/MTD/Customized_host.sh -t 20 -d http://ftp.ensembl.org/pub/release-104/fasta/callithrix_jacchus/dna/Callithrix_jacchus.ASM275486v1.dna.toplevel.fa.gz -c 9483 -g http://ftp.ensembl.org/pub/release-104/gtf/callithrix_jacchus/Callithrix_jacchus.ASM275486v1.104.gtf.gz`

- Users can update microbiome databases easily by one command line: **bash Update.sh -t [threads]**

  For example:

  `bash ~/MTD/Update.sh -t 20`

- Users can modify the contaminant list (conta_ls.txt) in the MTD folder by adding the taxonomy ID of the microbe in the second column of the list and its name in the first column (optional).

Licence