Meta-Transcriptome Detector (MTD): a novel pipeline for metatranscriptome analysis of bulk and single-cell RNAseq data

Fei Wu[1,2], Yaozhong Liu[3], Binhua Ling[1*]

1. Host-Pathogen Interaction Program, Texas Biomedical Research Institute, 8715 W Military Dr, San Antonio, TX 78227, USA
2. Tulane Center for Aging, Tulane University School of Medicine, New Orleans, LA 70112, USA
3. Tulane University School of Public Health and Tropical Medicine, New Orleans, LA 70112, USA

* Corresponding author: Binhua Ling, bling@txbiomed.org

Abstract

RNA-seq data contains not only host transcriptomes but also non-host information that comprises transcripts from active microbiota in the host cells. Therefore, metatranscriptomics can reveal gene expression of the entire microbial community in a given sample. However, there is no single tool that can simultaneously analyze host-microbiota interactions and to quantify microbiome at the single-cell level, particularly for users with limited expertise of bioinformatics. Here, we developed a novel software program that can comprehensively and synergistically analyze gene expression of the host and microbiome as well as their association using bulk and single-cell RNA-seq data. Our pipeline, named Meta-Transcriptome Detector (MTD), can identify and quantify microbiome extensively, including viruses, bacteria, protozoa, fungi, plasmids, and vectors. MTD is easy to install and is user-friendly. This novel software program empowers researchers to study the interactions between microbiota and the host by analyzing gene expressions and pathways, which provides further insights into host responses to microorganisms.

Key words: software, bulk RNAseq, single-cell RNAseq, metatranscriptomics, microbiota

1 Introduction

A variety of microorganisms have been recognized to contribute to the development of human diseases, including cancer, autoimmune diseases, and psychological disorders. For example, Helicobacter pylori can cause stomach cancer [1], human papillomavirus (HPV) infection can lead to uterine cervix cancer [2], and HIV-1 infection can lead to acquired immunodeficiency syndrome (AIDS) [3] as well as HIV-Associated Neurocognitive Disorders (HANDs) [4]. In addition, the Epstein–Barr virus (EBV) was found to contribute to 1.5% of total human cancers of various types worldwide [5]. However, systemic and comprehensive investigation of microorganisms in tissues and their contribution to disease development has yet to be conducted. In particular, the pathogenic mechanisms of a large number of opportunistic infections remain unexplored. Furthermore, different cell populations may contribute to the heterogeneous tropism in infections. Therefore, it is important to analyze microbiome diversity, abundance, their interaction with host cells, and impacts on infected cells.

Transcriptome from a host tissue may contain mRNAs from microorganisms that have not been fully investigated. Several tools have been developed to detect the microbiomes in the RNA-seq data, such as Kraken2 [6], VIRTUS [7], and IDseq [8]. However, it is lack of a software that can analysis both host and microbiome transcriptome in the same set of data, and it is also challenging for a researcher without bioinformatics expertise to examine microbiome in host tissues and its relation to the endogenous expression of host genes, especially at the single-cell level. To facilitate the effort of analysis of host transcriptome with its microbiome, we developed the Meta-Transcriptome Detector (MTD), a user-friendly pipeline for comprehensive and integrative investigation of microbiome from bulk and single-cell RNA-seq data.

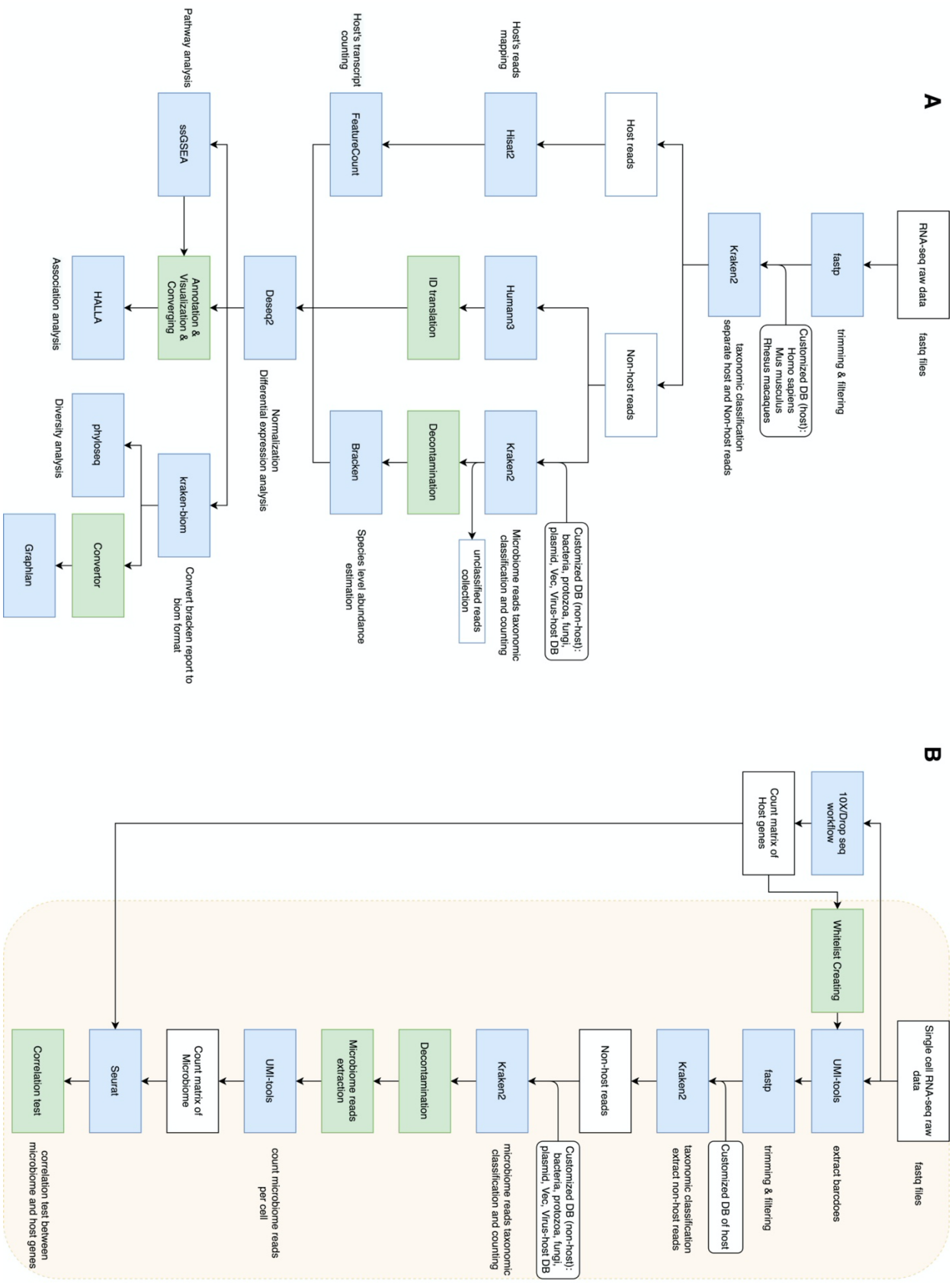


Figure 1. An overview of MTD. (A): A workflow diagram of MTD for bulk mRNA-seq analysis. (B): A workflow diagram of MTD for single-cell mRNA-seq analysis. The white boxes represent the reads in FASTQ format and the count matrix, the blue boxes show the bioinformatics software used, and the green boxes are the additional tools for data processing. The white boxes with curved edges show the reference genome and databases. In the single-cell mRNA-seq workflow (B), the left side exemplifies the host reads process protocols, and the right side in yellow shadow shows the MTD automatic pipeline to calculate the count matrix for the microbiome reads and the correlation test between microbiome and host genes.

2 Methods

2.1 Description of MTD

The MTD has two sub-pipelines to detect and quantify microbiomes by analyzing bulk and single-cell RNA-seq data, respectively (Figure 1). MTD is written in R (version 4.0.3) and Bash (version 4.2) languages and executed 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. The user manual for detailed instruction of installation and usage was disclosed on the webpage https://github.com/FEI38750/MTD. Here we describe the two sub-pipelines separately.

2.1.1 Transcriptome and meta-transcriptome analysis of bulk RNA-seq data

First, RNA-seq raw reads in the FASTQ file are trimmed and filtered by fastp (version 0.20.1) [9] with polyA/T trimming, and reads shorter than 40 bp (with the option --trim_poly_x --length_required 40) are discarded. Then, processed reads are classified based on the host genome by Kraken2 (version 2.1.1) [6] with default parameters. Finally, the host and non-host reads are organized separately in FASTQ format.

Host transcriptome analysis: The host species supported by MTD initially are Homo sapiens (Reference genome assembly: GRCh38), Mus musculus (Reference genome assembly: GRCm39), and Macaca mulatta (Reference genome assembly: Mmul_10). Additionally, users can add other host species by one command line. The host reads are aligned to the reference host

genome by Hisat2 (version 2.2.1) [10] with default parameters written in a SAM file. Quantification of reads for host gene expression is done by featureCounts (version 2.0.1) [11]. Next, the count data is analyzed by DESeq2 package (version 1.32.0) [12] in Bioconductor to get the differentially expressed genes (DEGs). The gene annotation is done through the biomaRt R package (version 2.46.3) [13, 14] in Bioconductor. The data visualization and a count matrix are automatically generated through R programs. The data visualization includes the heatmap, Venn Diagram, PCA, barplot, and volcano plot. The count matrix contains the Ensembl gene ID, gene symbol, chromosome name, gene position, functional descriptions, DEG results for each pairwise group comparison, raw, normalized and transformed reads counts. This count matrix is saved in CSV format, ready for downstream analyses such as pathway enrichment and customized data visualization.

Microbiome transcriptome analysis: MTD supports a broad spectrum of microbiome species and vectors including viruses, bacteria, protozoa, fungi, plasmids, and vectors. At the time of writing, the viruses contain 16,275 species from Virus-Host DB [15] and also simian immunodeficiency virus (SIVmac239) (GenBank accession number M33262). The rest of the microbiome are from the NCBI RefSeq database [16], which includes 63,237 species of bacteria, 13,970 fungi, 1,337 archaea, 573 protozoa, and 5,855 plasmids. In addition, vector contamination can be screened using the NCBI UniVec Database. Users can update the microbiome databases in MTD by one command line. The non-host reads are further classified by Kraken2 based on microbiome references with the default parameters, followed by a decontamination step that removes the microorganism under the genera reported as reagent and laboratory contaminants [17]. Users can easily customize the blacklist of contaminants based on their experimental environments. Then, the abundance of the microbiome in the species level is calculated by Bracken (version 2.6.0)

[18]. Next, the count data is imported into DESeq2 for analysis of the differentially expressed species. In DESeq2, we adjust the abundance of microbiome species based on the transcriptome size of the sample. The rationale is that count from a microbial species should take into account the overall representation of the host transcriptome. This normalization step is conducted through the formula, design ~ group + transcriptome_size, where transcriptome_size is defined by the formula: log2 (of a transcriptome size) - mean (of all log2-transformed transcriptome sizes in a sample), and group is the group code (e.g., A or B) of a sample. As a result, the importance of a microbial species in a sample with a higher transcriptome representation is underscored. Meanwhile, the non-host reads (including both unclassified and classified reads by Kraken2 using the reference databases) are imported into the Humann3 (version 3.0) [19] for profiling microbial metabolic pathways and molecular functions. The ChocoPhlAn [20] and full UniRef90 [21] are used as reference databases for nucleotide and protein, respectively. Then the profiling results are annotated to understandable functional terms to facilitate the downstream analyses. Next, the heatmaps of DEGs, Venn Diagrams, PCA, bar plots, and volcano plots are generated for the microbial species, molecules, and metabolic pathways based on the Deseq2 results. Additionally, kraken-biom (version 1.0.1) [22] is used to format the data for diversity analyses and phylogenetic tree plotting. The phyloseq R package (version 1.34.0) [23] in Bioconductor and vegan R package (version 2.5-7) [24] are used to analyze the diversities, including alpha diversity (Shannon, Simpson) and beta diversity (Bray-Curtis). Then, box plots are shown to visualize the t-test comparison results between groups, including alpha diversities of classified reads and the abundances of unclassified reads. PCoA graphing and analysis of similarities (ANOSIM) are based on Bray-Curtis distance. The relative abundance of a microbial species in the total microbiome is shown as bar plots at the Phylum level. The heatmap of total microbiome abundance is plotted

using data normalized by Deseq2. Next, the phylogenetic trees are plotted through modified Graphlan [25], which is a tool for generating informative and integrative circular graph representing phylogenetic and taxonomic trees. However, the original program requires a specific data format as input and is not compatible with the output from Kraken2 and Bracken software. Therefore, we wrote a converter program and integrated it into the pipeline to bridge the taxonomic classification software Kraken2/Bracken and graph-making software GraPhlAn. It allows us to transform the output of Kraken2 and Bracken to match the data structure requirement of GraPhlAn. Furthermore, the default settings of colors have been optimized by modifying the source code of GraPhlAn. In addition to .biom format, data is also saved in .mpa and .krona formats to facilitate further downstream visualizations.

Most importantly, in the final steps of the pipeline, MTD examines the association between the microbiome and the host's characteristics, such as gene expression and pathways. Pathway enrichment for each sample is performed by the single sample Gene Set Enrichment Analysis (ssGSEA 2.0) program [26-28] with a MSigDB C2 database that contains 6,290 curated gene sets. ssGSEA 2.0 is modified for parallel computing in High-Performance Computing (HPC) environment, which is described in the supplementary document. Next, the effects from covariates among groups are adjusted through the removeBatchEffect function in the limma R package (version 3.48.1) [29] in Bioconductor. The association analyses are then conducted through Halla (version 0.8.18) [30], which is set to compute hierarchical clustering of Spearman pairwise correlation. Figure 2 illustrates the MTD automatic pipeline for dual-analyzing the bulk RNA-seq raw data.

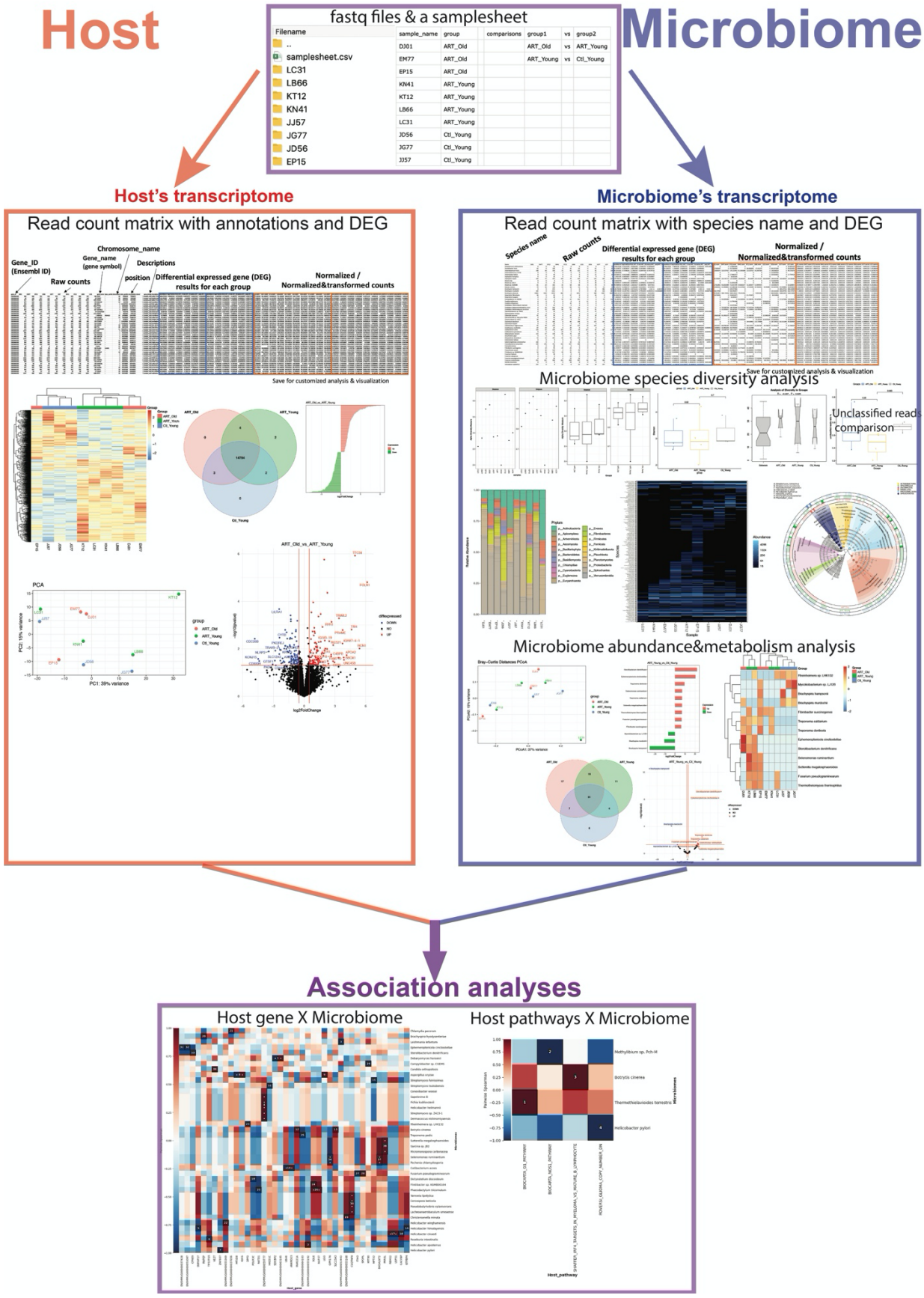


Figure 2. The automatic pipeline of MTD for dual analysis of the bulk RNA-seq raw data. Analysis results are automatically saved into the folder assigned by the user. Examples of analysis outputs for the transcriptome of host and microbiome are demonstrated on the left side and right side, separately. The shared procedures are shown in the purple boxes, which include the input files (upper box) and the association analyses (bottom box). In addition to the graphs, all the detailed information was included in the data sheets and stored in the corresponding output folder. To use the MTD, the user must simply put the FASTQ files in a folder with a sample sheet in CSV format that describes the sample names, groups, and comparisons, and then perform the analysis with one command line. For example, with the command line "bash MTD.sh -i ~/inputpath/samplesheet.csv -o ~/outputpath -h 9544 -t 20", the user enters the place of the sample sheet with the raw data after the flag "-i", and where one wants to place the results (after flag "-o"), the host taxonomic ID is after "-h", and the threads of CPU after "-t".

2.1.2 Analysis of single-cell RNA-seq data

The MTD supports automatic generation of the count matrix of the microbiome by using raw data in the FASTQ format and count matrix of host genes from two commonly used single-cell RNA-seq platforms, 10x Genomics and Drop-seq. First, the user can download the available count matrix of host genes (H5, Matrix, or .dge.txt format) or follow the corresponding workflow to process the raw data, such as through the 10x Cell Ranger software. The cell barcodes are identified from the count matrix of host genes, then the UMI and cell barcodes are extracted and added to read names by using the UMI-tools (version 1.1.1) [31]. Second, the reads are trimmed and filtered by fastp, then filtered for the host reads by Kraken2. The non-host reads are further classified by Kraken2 with the comprehensive microbiome databases, followed by a decontamination step. The steps from trimming to decontamination use the same settings as mentioned in the previous section. Next, the taxonomic labels of the reads are extracted and aligned with corresponding cell barcodes through a step written by the AWK program language. Finally, UMI-tools is used to generate a count matrix (tab-separated) through a converter written in R. Figure 3 exemplifies a count matrix automatically generated by MTD for the microbiome in each cell. Subsequently, MTD combines the count matrices of the host genes and microbiome to perform the correlation analysis automatically. An example of the correlation analysis result is showed in Figure 7 (A-B).

At the single-cell level, the Spearman correlations between microbial organisms and host genes are tested by using the top 3,000 most highly variable features, including the normalized data of both host and microbiome. Because the step of correlation analysis step is highly time-consuming for a large data matrix, parallelizing computing was applied to speed up the computation by using the doParallel R package (Version 1.0.16). The other analysis methods through Seurat R package (Version 4.0.1) and homemade programs are described in the supplementary document. The diagram of the MTD pipeline for single-cell RNA-seq is demonstrated in Figure1 B.

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

Figure 3. An example of the count matrix that is automatically generated by MTD for the single-cell microbiome analysis. For illustration, the figure shows part of the large count matrix. The name and taxonomy ID of the microbiome is shown in the first column and highlighted in the green box. The read counts are highlighted in the blue box. The first row shows the cell barcodes.

2.2 Animal information and sample collection

Here we demonstrated the application of MTD for analyzing the bulk RNA-seq data by using samples from the descending colon and brain mononuclear cell (BMC) from rhesus monkeys as an example. The detailed sample information is in the supplementary document.

Animal euthanasia was performed in line with the recommendations of the Panel on Euthanasia of the American Veterinary Medical Association. Following Tulane IACUC standards

of operation (SOP), SIV-infected and/or drug-treated macaques were euthanized first with telazol and buprenorphine, followed by sodium pentobarbital intravenous injection. Tissues from descending colon were collected fresh and soaked in RNAlater™ Stabilization Solution during necropsies. To avoid the chance of microbiome contamination from the animal gut to the brain, brain tissues were collected first, and the descending colons were collected last. After BMC isolation, cells were sorted into high autofluorescent (BMC_H) and low autofluorescent (BMC_L) groups, based on cellular autofluorsecent levels. The methods of BMC isolation, cell sorting, and the bulk RNA-seq protocol are described in the supplementary document.

Raw data of the singe-cell RNA-seq was downloaded from NCBI GEO with accession code GSE161340 (SRR13041553-13041560), and GSE160384 (SRR12933210-SRR12933217). The detailed sample preparation methods for the sequencing were described in the articles [32, 33]. The detailed methods for single-cell RNA-seq analysis are in the description in supplementary document.

3 Results

3.1 Application to bulk RNA-seq analysis: descending colon of rhesus macaques

Through MTD, the transcripts of both the microbiome and the host were analyzed simultaneously using the same bulk RNA-seq raw data. Figure 4 is a heatmap showing the abundance of all the microbiome species in the samples from the descending colon. Supplementary Figure 1 presents the taxonomic and phylogenetic trees of microorganisms detected in the descending colon samples from rhesus macaques. More detailed results of the microbiome and host gene analyses of the descending colon are described in the supplementary document sections 2.1 and 2.2, respectively. The microbiome analysis results of BMC are in supplementary document sections 2.3. Here we demonstrate the results of association analyses of the microbiome to the host genes or pathways in descending colon.

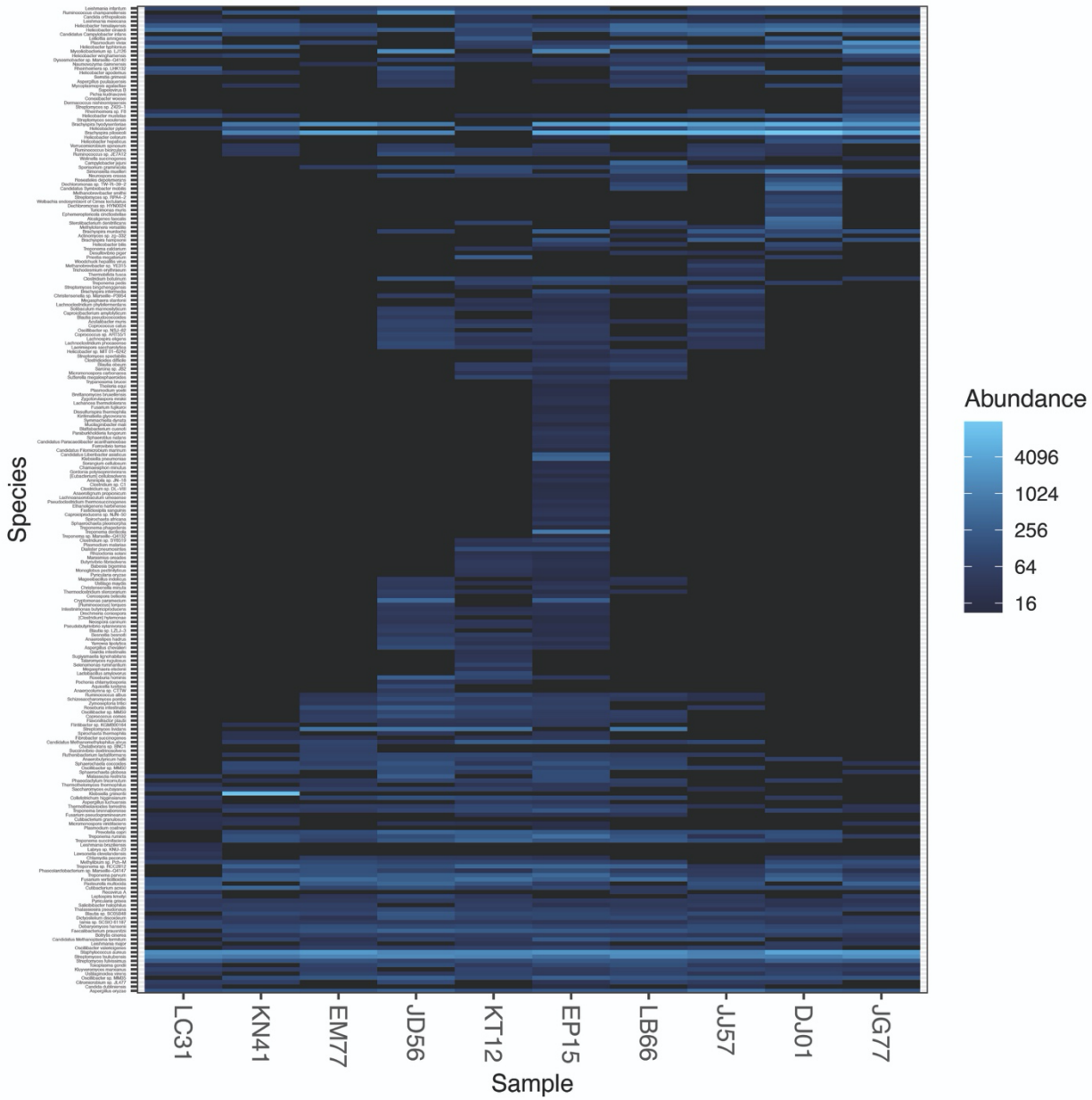


Figure 4. Heatmap of the microbiome species identified in descending colon samples of rhesus macaques. The higher abundance of microbiome reads is shown in a light blue color, and the lower abundance is shown in a darker color. Data was normalized using the Deseq2 and plotted by the phyloseq R packages, which were wrapped in MTD.

The correlations between the microbiome and host gene are illustrated in Figure 5A. For example, the *Debaryomyces hansenii* shows a significant positive correlation with host gene *UBE2I* and *IL27RA*. The expression of the host gene *C1QTNF8* positively correlated with a group

of microbes, such as *Yarrowia lipolytica*, *Aspergillus chevalieri*, *Roseburia hominis*, and *Anaerostipes hadrus*.

The correlations between the microbiome and host pathways are shown in Figure 5B. For example, the pathway that controls the amplification of the 8q24 chromosome region (HEIDENBLAD_AMPLICON_8Q24_UP) was upregulated with the mRNA expression of *Saccharomyces eubayanus*. The complex I biogenesis signaling pathway (REACTOME_COMPLEX_I_BIOGENESIS) was negatively correlated with the expression of *Helicobacter cinaedi*.

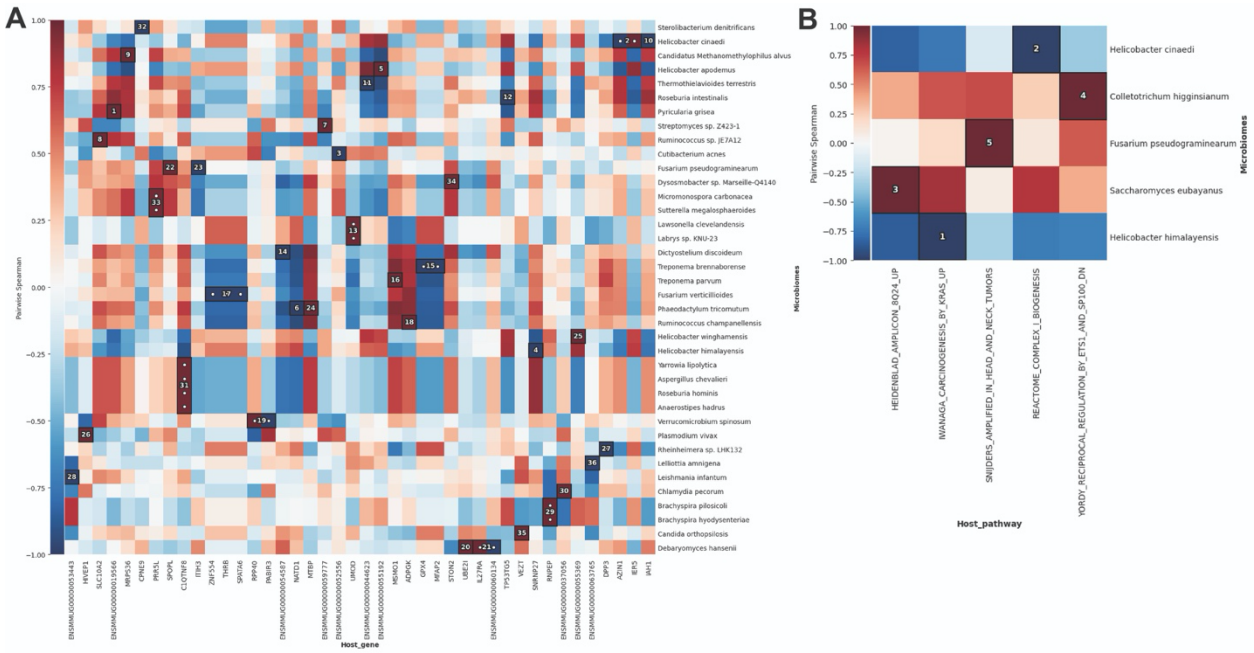


Figure 5. Analyses of association between the microbiome and host genes or pathways in descending colon of the rhesus macaque. The figure shows the correlation between the RNA expression level of microbiome species and host gene (A), or pathways (B). The x-axis is labeled with the names of the host genes or pathways, and the y-axis lists the names of microbiome species. Positive correlation coefficients are shown in red, and negative correlation coefficients are shown in blue. The significant results are marked with white dots and ranked by numbers. The results in the same cluster can be found in a box with the same number. The association was examined by pairwise Spearman correlation test based on hierarchical clustering. The tabulates of all comparison results and dot plots were saved in the corresponding output folder.

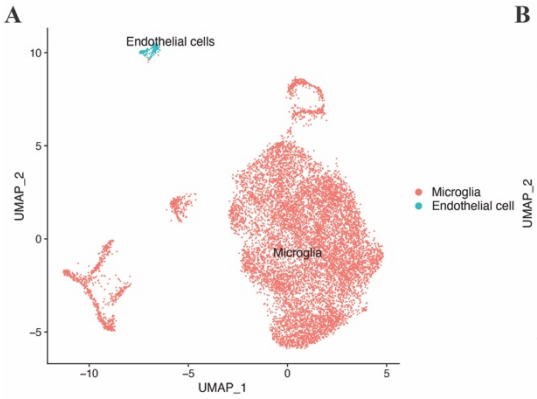
3.2 Application to single-cell RNA-seq analysis

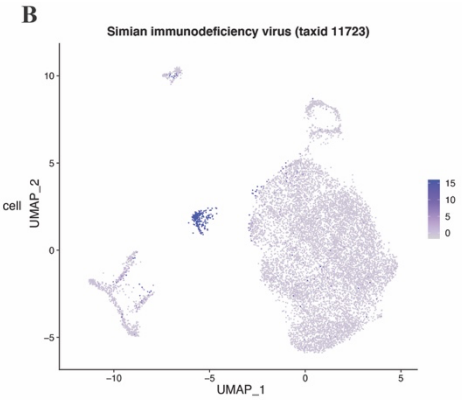
3.2.1 Microglia cells of SIV-infected rhesus macaques

We next applied MTD to single-cell RNA-seq data from microglia cells isolated from SIV-infected rhesus macaques [32]. Because the analysis results from the authors identified SIV transcripts in the single cells, it is an ideal dataset for validation of the capacity of MTD to process single-cell RNA-seq data.

First, count matrices of the microbiome were generated by MTD. Then, they were integrated with the host transcriptome for downstream analysis through the Seurat R package. The results showed that the major cell type in the sample was microglia, with a small portion of endothelial cells (Figure 6A). A cluster of microglia cells was highly infected with SIV (Figure 6B). For this cluster, the top 20 makers ranked by folder change are displayed in Figure 6C, such as PDE4A and SENP3. The cell subpopulation with these markers implicated a higher SIV tropism.

Overall, the results validated the capability of MTD for detecting specific microorganism species from single-cell RNA-seq data.





C

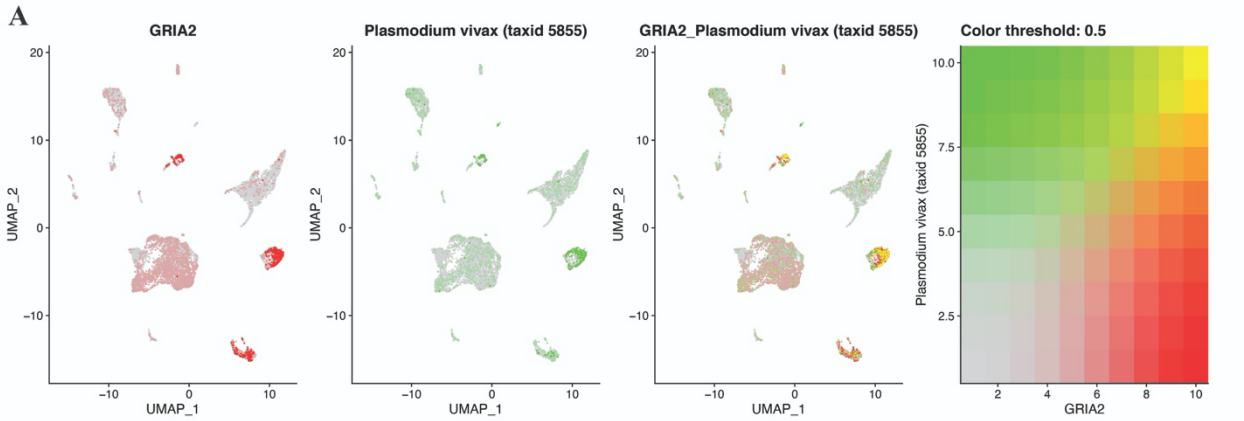
	avg_log2FC	p_val_adj
PDE4A	12.5733556	1.85E-47
SENP3	11.4887558	6.01E-63
LOC114673080	6.47664879	0.00060731
LOC106993196	5.40468811	7.42E-07
SERPING1	5.37523375	9.29E-14
LOC711690	4.77859188	1.24E-49
CLEC6A	4.6409414	8.13E-76
HIST1H2BH	4.62930683	5.65E-36
CYB5RL	4.39356614	3.91E-43
SLC29A3	4.17949231	8.05E-19
MOCS3	4.13607255	1.16E-52
SIX3	4.02465922	5.86E-128
TNRC6C	3.95155637	8.70E-38
SEMA7A	3.68049944	1.74E-07
ZNF345	3.62944169	1.66E-54
KIF21A	3.52219847	1.24E-39
OASL	3.51677381	0.01163059
LOC106998309	3.4349693	8.69E-05
UACA	3.37391018	1.33E-74
TRIM44	3.30532868	0.00581639

Figure 6. Detection of SIV in microglia cells from rhesus macaques. (A) Cell types in different colors on the UMAP plot. Cell type identity was assigned based on the homemade program described in supplementary document. (B) SIV reads detected by MTD. The blue dots on the UMAP plot indicate the SIV-infected cells with the normalized reads quantity. (C) Markers of the cell cluster that harbor SIV. Analyses of the count matrix followed by visualization were performed through Seurat. Each dot in the UMAP plot represents every single cell. FindMarkers function with MAST methodology was used for computing the log2fold changes for each gene/feature between clusters and their corresponding adjusted p-values.

3.3.2 Brain cells of mice

We also applied MTD on the single-cell RNA-seq data of brain cells isolated from mice [33] and demonstrated the correlation analyses between the microbiome and host genes or pathways at the single-cell level.

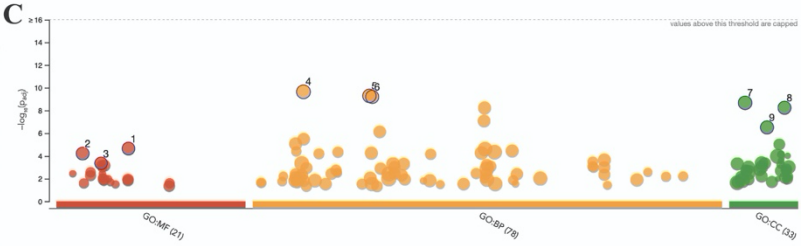
We found that *plasmodium vivax* (*P. vivax*) and host gene *GRIA2* had the highest correlation coefficient (Figure 7A). As shown in Figure 7B, *GRIA2* was highly expressed in the *P. vivax*-infected cells. We further identified all the host genes that highly correlated with *P. vivax* (Figure 7C), then performed pathway enrichment analysis. The results underscored the positive association of *P. vivax* with the function of the infected cell's plasma membrane region, such as cell junction and transmembrane transporter activity (Figure 7D). This result supports the cytoadherence phenomenon of *P. vivax* reported in previous research [34-36]. Although *P. vivax* primary infects red blood cells, its cytoadherence on other cell types has been reported, such as in endothelial cells [34, 35]. Moreover, recent findings suggest that it has the ability to adhere to all Chinese hamster ovary (CHO) cells [35]. Our results brought insights into the interaction between *P. vivax* and host cells. It showed that *P. vivax* interacts with host cells that are incrementally expressing genes of the cellular membrane. Future research can study the causal effect of these molecules during infection, such as whether they contribute to pathogen adherence or if the infection leads to their increased expression.



B

Microbiome	Host_gene	r	p
Plasmodium vivax (taxid 5855)	GRIA2	0.28639678	1.60E-133
Hafnia alvei (taxid 569)	TNPO3	0.27675073	1.87E-124
Hafnia alvei (taxid 569)	CITED2	0.27392659	7.22E-122
Plasmodium vivax (taxid 5855)	NTM	0.27211198	3.19E-120
Plasmodium malariae (taxid 5858)	NTM	0.2667293	2.05E-115
Plasmodium vivax (taxid 5855)	GPM6A	0.2600528	1.32E-109
Plasmodium vivax (taxid 5855)	GJA1	0.25837128	3.61E-108
Plasmodium malariae (taxid 5858)	RORB	0.25833362	3.89E-108
Plasmodium vivax (taxid 5855)	LSAMP	0.25418133	1.23E-104
Plasmodium malariae (taxid 5858)	PLPP3	0.25278256	1.81E-103
Plasmodium vivax (taxid 5855)	SLC4A4	0.25264053	2.37E-103
Plasmodium vivax (taxid 5855)	DCLK1	0.25165597	1.55E-102
Plasmodium malariae (taxid 5858)	DCLK1	0.25043666	1.57E-101
Plasmodium vivax (taxid 5855)	NTRK2	0.24976506	5.57E-101
Hafnia alvei (taxid 569)	EGR1	0.24960357	7.55E-101
Plasmodium vivax (taxid 5855)	NKAIN4	0.24871877	3.99E-100
Hafnia alvei (taxid 569)	NFKBIZ	0.24835107	7.95E-100
Hafnia alvei (taxid 569)	NFKBIA	0.24741651	4.56E-99
Plasmodium malariae (taxid 5858)	SLC4A4	0.24660404	2.07E-98
Plasmodium vivax (taxid 5855)	TMEM47	0.24635789	3.28E-98

C



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0022857	transmembrane transporter activity	2.262×10^{-5}
2	GO:MF	GO:0005215	transporter activity	6.178×10^{-5}
3	GO:MF	GO:0015075	ion transmembrane transporter activity	4.414×10^{-4}
4	GO:BP	GO:0007399	nervous system development	2.281×10^{-10}
5	GO:BP	GO:0022008	neurogenesis	5.183×10^{-10}
6	GO:BP	GO:0030182	neuron differentiation	6.301×10^{-10}
7	GO:CC	GO:0030054	cell junction	2.126×10^{-9}
8	GO:CC	GO:0098590	plasma membrane region	5.753×10^{-9}
9	GO:CC	GO:0045202	synapse	3.097×10^{-7}

Figure 7. Co-expression microbiome and host genes in host cells. (A) Visualization of the highest correlation on the UMAP plot. The cells expressed GRIA2 are represented by red dots, and the cells infected with plasmodium vivax are shown as green dots. The cells containing both are represented by the overlapping of the two colors (yellow). (B) A list of the top 20 correlations between host genes and microorganisms, and the highest pair is highlighted in red. (C) The results of pathway enrichment of the genes that were highly associated with *plasmodium vivax*, which were defined by $r > 0.2$ and $p < 0.05$. The top 3 results of each GO categories are highlighted. UMAP plots were drawn by the Seurat R package. The pathway enrichment was performed through g:GOSt in g:Profiler (version: e104_eg51_p15_3922dba, organism: mmusculus).

4 Conclusion

MTD is a novel metatranscriptome detection and quantification pipeline, which can perform both bulk and single-cell RNA-seq data analysis. With this software, the activated microbiome (including the virome) can be detected in the sample, the cell type harboring them can be identified, and the host gene expression and microbial prevalence can be correlated. Thus, it would be a useful tool to improve our understanding of host-pathogen interactions, in particular, how the microbiome contributes to the host health, and what genes and pathways of the host are important to a particular infection by a microbial species, which may shed lights to treatment and prevention of infectious diseases.

5 Discussion

MTD has several unique advantages compared to current tools in detecting meta-transcriptome. MTD can simultaneously and comprehensively profile microbiome and the host transcriptome in both bulk and single-cell RNA-seq data. The program takes into account the host transcriptome size while modeling microbiome reads in differential abundance analysis. In addition, our program has a decontamination step to eliminate the potential noise from the contaminant microbes. Specifically, MTD has a decontamination step that blacklists the common contaminant microbes in the laboratory environment [17]. Users can modify the list depending on the contaminant in their situations. Moreover, MTD warrants better accessibility as well as data safety. The software installation, updating, and transcriptomic analyses can all be performed by a single command. It obviates the need of using other cloud-based applications.

Currently, it is common to perform the polyA tail enrichment during the library preparation for mRNA sequencing. Thus, MTD could avoid contamination from viruses in most cases because virus RNA only acquires polyA tail when it is transcribed in the host cell. Nevertheless, users need to be cautious about the exceptions of single-strand RNA viruses and other preparation methods that contain the polyA tail. Moreover, it is still challenging to detect and remove other contaminated microbes. There are tools to identify the potential contaminant by simply calculating the correlation of nucleic acid abundance between microbes and host [37]. However, these methods may not be sensitive if the samples have similar abundance, heterogeneous contaminant patterns, or cross-contamination.

In the era of single-cell genomics consortia, the distribution of the microbiome in the cell population, tissues, and organ levels, and their associations with cell functions will be better analyzed and further understood. Researchers will be able to get insight into the pathogenesis of

each microorganism identified. Furthermore, annotating the sample's geographic information with each microorganism would offer us a map of pathogens which could predict an epidemic. Thus, MTD could become a critical element for monitoring the spread of the microbiome and its pathogenesis in the future.

Key Points

- MTD enables simultaneous detection of the microbiome in the cells and the host gene expression in bulk and single-cell RNA-seq data.
- The association between the microbiome and the host gene expression or pathways is automatically analyzed.
- MTD has an extensive microbiome detection ability, including viruses, bacteria, protozoa, fungi, plasmids, and vectors.
- Install and use MTD by one command line without the requirement of administrator/root privilege.
- Decontamination function to eliminate the common contaminant microbes in the laboratory environment.

Supplementary Data

Supplementary data are available online at *bioRxiv*.

Data Availability

MTD software can be accessed through GitHub at https://github.com/FEI38750/MTD

Funding

This work was supported by National Institutes of Health R01 MH116844 (BL) and R01 NS104016 (BL). The funders had no role in study design, data collection and analysis, preparation of the manuscript or decision for publication.

Acknowledgements

The authors thank Sandra Smith and Brian Kopecki at the Department of Information Technology of Texas Biomedical Research Institute to maintain HPC and install the third-party programs asked. We are grateful to any contributors in our respectable open-source community and their virtue in sharing the questions and answers.

References

1. Wroblewski, L.E., R.M. Peek, Jr., and K.T. Wilson, *Helicobacter pylori and gastric cancer: factors that modulate disease risk.* Clinical microbiology reviews, 2010. **23**(4): p. 713-739.
2. Muñoz, N., et al., *Chapter 1: HPV in the etiology of human cancer.* Vaccine, 2006. **24**: p. S1-S10.
3. Sharp, P.M. and B.H. Hahn, *Origins of HIV and the AIDS pandemic.* Cold Spring Harbor perspectives in medicine, 2011. **1**(1): p. a006841-a006841.
4. Cohen, R.A., T.R. Seider, and B. Navia, *HIV effects on age-associated neurocognitive dysfunction: premature cognitive aging or neurodegenerative disease?* Alzheimer's Research & Therapy, 2015. **7**(1): p. 37.
5. Farrell, P.J., *Epstein–Barr Virus and Cancer.* Annual Review of Pathology: Mechanisms of Disease, 2019. **14**(1): p. 29-53.
6. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biol, 2019. **20**(1): p. 257.
7. Yasumizu, Y., et al., *VIRTUS: a pipeline for comprehensive virus analysis from conventional RNA-seq data.* Bioinformatics, 2020.
8. Kalantar, K.L., et al., *IDseq-An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring.* Gigascience, 2020. **9**(10).
9. Chen, S., et al., *fastp: an ultra-fast all-in-one FASTQ preprocessor.* Bioinformatics, 2018. **34**(17): p. i884-i890.
10. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature Biotechnology, 2019. **37**(8): p. 907-915.
11. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2013. **30**(7): p. 923-930.
12. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
13. Durinck, S., et al., *BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.* Bioinformatics, 2005. **21**(16): p. 3439-3440.
14. Durinck, S., et al., *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.* Nature Protocols, 2009. **4**(8): p. 1184-1191.
15. Mihara, T., et al., *Linking Virus Genomes with Host Taxonomy.* Viruses, 2016. **8**(3): p. 66.
16. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic acids research, 2005. **33**(Database issue): p. D501-D504.
17. Salter, S.J., et al., *Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.* BMC Biology, 2014. **12**(1): p. 87.
18. Lu, J., et al., *Bracken: estimating species abundance in metagenomics data.* PeerJ Computer Science, 2017. **3**: p. e104.
19. Beghini, F., et al., *Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.* eLife, 2021. **10**: p. e65088.
20. Franzosa, E.A., et al., *Species-level functional profiling of metagenomes and metatranscriptomes.* Nat Methods, 2018. **15**(11): p. 962-968.
21. Suzek, B.E., et al., *UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches.* Bioinformatics, 2014. **31**(6): p. 926-932.
22. Dabdoub, S., et al., *kraken-biom*. 2016: GitHub.

23. McMurdie, P.J. and S. Holmes, *phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.* PLoS One, 2013. **8**(4): p. e61217.
24. Oksanen, J., et al., *vegan: Community Ecology Package*. 2020.
25. Francesco Asnicar, et al., *GraPhlAn*. 2020.
26. Barbie, D.A., et al., *Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1.* Nature, 2009. **462**(7269): p. 108-112.
27. Krug, K., et al., *A Curated Resource for Phosphosite-specific Signature Analysis.* Molecular & cellular proteomics : MCP, 2019. **18**(3): p. 576-593.
28. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
29. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015. **43**(7): p. e47-e47.
30. Gholamali Rahnavard, et al., *High-sensitivity pattern discovery in large multi'omic datasets*. 2021.
31. Smith, T., A. Heger, and I. Sudbery, *UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy.* Genome research, 2017. **27**(3): p. 491-499.
32. Niu, M., et al., *Methamphetamine Increases the Proportion of SIV-Infected Microglia/Macrophages, Alters Metabolic Pathways, and Elevates Cell Death Pathways: A Single-Cell Analysis.* Viruses, 2020. **12**(11).
33. Ogrodnik, M., et al., *Whole-body senescent cell clearance alleviates age-related brain inflammation and cognitive impairment in mice.* Aging Cell, 2021.
34. De las Salas, B., et al., *Adherence to human lung microvascular endothelial cells (HMVEC-L) of Plasmodium vivax isolates from Colombia.* Malar J, 2013. **12**(1): p. 347.
35. Salazar Alvarez, L.C., et al., *Plasmodium vivax Gametocytes Adherence to Bone Marrow Endothelial Cells.* Front Cell Infect Microbiol, 2021. **11**: p. 614985.
36. Totino, P.R. and S.C. Lopes, *Insights into the Cytoadherence Phenomenon of Plasmodium vivax: The Putative Role of Phosphatidylserine.* Frontiers in Immunology, 2017. **8**(1148).
37. Davis, N.M., et al., *Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data.* Microbiome, 2018. **6**(1): p. 226.