BMC Bioinformatics

RESEARCH

Open Access

metaTP: a meta-transcriptome data analysis pipeline with integrated automated workflows

Limuxuan He[1], Quan Zou[1,2] and Yansu Wang[1*]

*Correspondence:
wangyansu@uestc.edu.cn

[1] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China
[2] Macao Polytechnic University, Macau Peninsula Gomes Street, Macau, 999078, China

Abstract

Background: The accessibility of sequencing technologies has enabled meta-transcriptomic studies to provide a deeper understanding of microbial ecology at the transcriptional level. Analyzing omics data involves multiple steps that require the use of various bioinformatics tools. With the increasing availability of public microbiome datasets, conducting meta-analyses can reveal new insights into microbiome activity. However, the reproducibility of data is often compromised due to variations in processing methods for sample omics data. Therefore, it is essential to develop efficient analytical workflows that ensure repeatability, reproducibility, and the traceability of results in microbiome research.

Results: We developed metaTP, a pipeline that integrates bioinformatics tools for analyzing meta-transcriptomic data comprehensively. The pipeline includes quality control, non-coding RNA removal, transcript expression quantification, differential gene expression analysis, functional annotation, and co-expression network analysis. To quantify mRNA expression, we rely on reference indexes built using protein-coding sequences, which help overcome the limitations of database analysis. Additionally, metaTP provides a function for calculating the topological properties of gene co-expression networks, offering an intuitive explanation for correlated gene sets in high-dimensional datasets. The use of metaTP is anticipated to support researchers in addressing microbiota-related biological inquiries and improving the accessibility and interpretation of microbiota RNA-Seq data.

Conclusions: We have created a conda package to integrate the tools into our pipeline, making it a flexible and versatile tool for handling meta-transcriptomic sequencing data. The metaTP pipeline is freely available at: https://github.com/nanbei45/metaTP.

Keywords: metaTP pipeline, Meta-transcriptome, Transcript expression quantification, Functional annotation, Co-expression network analysis

Background

Microbiome encompasses interacting communities of bacteria, fungi, archaea, protozoa, and viruses, which have been proven to directly and indirectly influence the health of plants and animals in various ecosystems [1]. The term 'meta-transcriptome' refers to the collective microbial RNA sequences, which can bridge metagenomic annotations

and facilitate a unified classification framework. Analyzing the meta-transcriptome can reveal variations in gene expression that are associated with specific phenotypes and provide insights into the active metabolic and functional roles of microbes. Unlike traditional transcriptomics, which focuses on the gene expression of a single species, metatranscriptomics investigates the transcriptional activities of all microorganisms within complex ecosystems. This broad approach makes it highly applicable in fields such as ecology, environmental science, and public health [2–4]. For example, the soil microbiome provides critical contributions to soil fertility, biogeochemical cycling, and plant health [5, 6], while the gut microbiota regulates intestinal and immunological homeostasis [7, 8], further demonstrates the profound impact of microbial community diversity and functionality on ecosystem dynamics.

Although metatranscriptomics has shown great potential in revealing microbial functions and metabolic activities, its data analysis still faces numerous challenges. Metatranscriptomic data, typically generated by high-throughput sequencing technologies, are characterized by large volumes, high noise levels, and the complexity of multispecies gene information, which pose significant difficulties in data processing and analysis. Similar to metagenomes, technical sequencing errors and other technical artefacts are commonly present in RNA-Seq technologies customarily [9, 10]. Given the large sample size required for analysis to reveal large-scale patterns, it requires pipelines that integrate multiple computational tools for robust meta-analysis, including data exploration, quality control, differential genes expression analysis, functional annotation, and further downstream analyses. The standard operating procedures have become essential for the scalability and reproducibility of data [11].

In recent years, several tools focusing on metatranscriptomic data analysis have been developed. However, the entire analysis process involves the installation and execution of various software and requires substantial computational resources. Web-based analysis platforms such as COMAN [12] and MG-RAST [13] are online tools dedicated to metatranscriptomic data analysis, but their analytical depth is limited, and their processing speed is constrained by server performance, which may result in suboptimal performance with large-scale datasets. IMP [14] provides a workflow that can handle both metagenomic and metatranscriptomic data, supporting reference genome-independent analysis methods, but its analysis speed is relatively slow, especially with complex samples, and it poses a learning curve for users without a bioinformatics background. HUMAnN2 [15] enables the integration of genomic and transcriptomic data for functional annotation and metabolic pathway analysis; however, it has high computational resource requirements and is slower in processing. While these tools perform well for specific tasks, they often lack systematic integration, requiring trade-offs based on the specific task requirements, user technical background, and available computational resources, making it difficult to meet the needs of end-to-end analysis. Therefore, there is an urgent need for an integrated, automated analysis pipeline that can efficiently process and comprehensively analyze data, simplifying the analysis process and improving the reliability and reproducibility of results.

To address the challenges, this study developed metaTP, an integrated automated pipeline designed to streamline the analysis of metatranscriptomic data. MetaTP seamlessly integrates modules for quality control, non-coding RNA removal, functional

annotation, and network analysis, providing a comprehensive, end-to-end automated pipeline from raw sequence data to downstream analysis. Compared to existing tools, metaTP provides superior efficiency and reproducibility.

Implementation

metaTP is a command line software for Unix-like systems that can execute seven steps. The metaTP pipeline consists of an independent program responsible for preprocessing the reads, and sequence processing (Fig. 1). Compared to existing methods that involve manual execution of data processing steps, metaTP achieves a high degree of automation in the data processing workflow by integrating the Snakemake workflow engine, thereby optimizing various stages of metatranscriptomic analysis. The pipeline accepts paired or single-end sequencing reads in FASTQ format as input through command line flags. It also provides progress reports and visualization options suitable for conducting exploratory factor analysis.

Data collection, quality control and non-coding RNA removal

The metaTP pipeline integrates data download options using the SRA Toolkit [16]. The target sequence run was downloaded in compressed sra format (.sra) using the

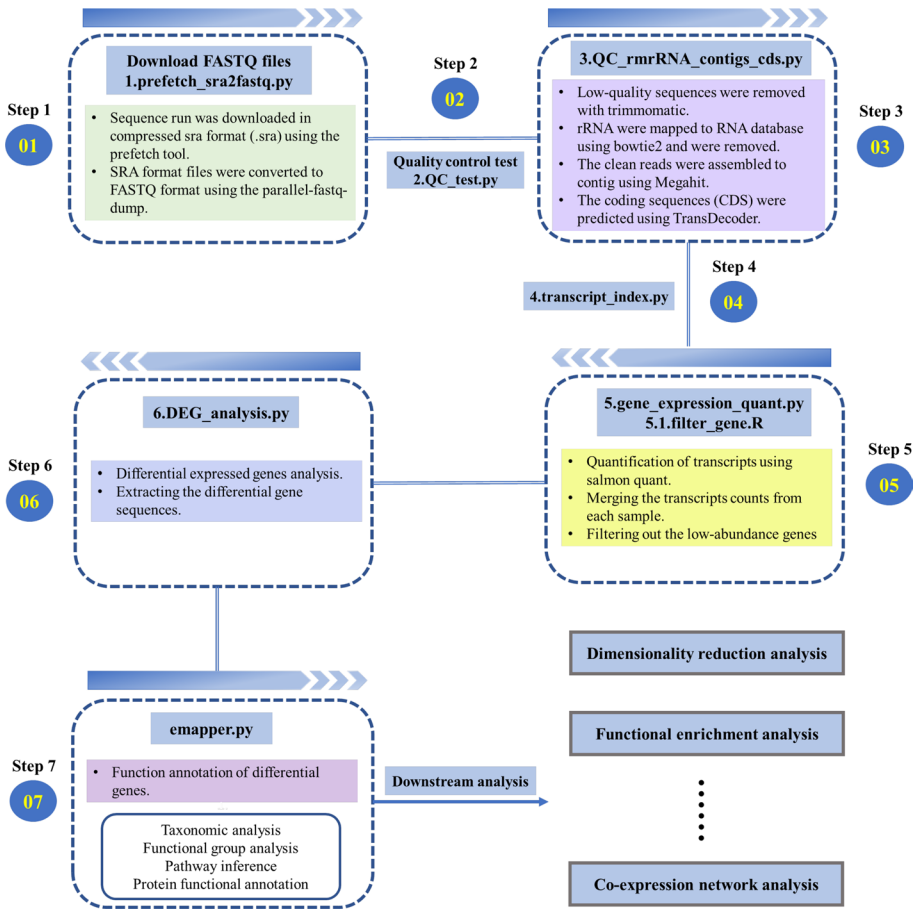

Fig. 1 Flow chart of metaTP

prefetch tool. The SRA files were then decompressed to fastq files using fasterq-dump. The quality of the FASTQ files is assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmomatic was employed to handle poor quality or technical sequences such as adapters [17]. The quality-trimmed fastq files were imported into bowtie2 for rRNA removal [18]. All the high-quality reads after depletion are further assembled using megahit, which makes use of succinct de Bruijn graphs [19]. Transdecoder was used to identify putative coding regions from the assembled transcripts and/or contigs based on the NCBI non-redundant (nr) protein database (http://transdecoder.github.io). Finally, the seqkit tool is used to deduplicate the predicted coding region sequences, generating a file free from redundant sequences.

Transcript expression quantification and differential gene expression analysis

MetaTP provides function Salmon for rapid transcript-level expression quantification. In meta-transcriptome analyses, traditional transcriptome quantitative analysis is not possible due to the absence of a reference genome annotation file (gtf/gff file). Salmon is the first transcriptome quantifier that corrects for GC fragment content bias, leading to improved accuracy of abundance estimates and reliability of subsequent differential expression analyses [20]. This makes it suitable for transcriptome and metagenomic analysis. To perform quantification, coding sequences obtained from Transdecoder are used to construct an index. A decoy aware transcriptome file is then created for Salmon transcript quantification, followed by the transcriptome index. Gene expression levels are normalized by transcript length and library size (TPM). Differential gene expression analysis in this study was performed using R and associated packages for data processing, with statistical testing of gene expression differences conducted based on the Wilcoxon rank-sum test. The analysis was carried out by filtering the expression values within the gene expression data matrix.

Functional annotation of genes and reads

Gene annotation was conducted using the metaTP that integrates eggNOG-mapper (http://eggnog-mapper.embl.de), a tool for functional annotation based on precomputed orthology assignments [21] and has been optimized for handling vast amounts of metagenomic data. It offers several key functionalities, including: (1) de novo gene prediction based on raw alignments, (2) integrated pairwise homology prediction, (3) rapid protein domain detection. The annotation results include gene functional categories, metabolic pathway information, GO annotations, and other relevant functional details. Additionally, functional enrichment analysis of the annotation results (such as GO and KEGG pathway analysis) is performed, and protein structure prediction is conducted to support the in-depth interpretation of the biological significance of differentially expressed genes.

Automation based on Snakemake execution

To enhance the automation level of the entire analysis workflow, metaTP utilizes Snakemake as the workflow engine to manage each step of data processing and analysis. Snakemake supports automated task execution, enabling tasks to be scheduled and processed in parallel based on dependency relationships, significantly improving analysis

efficiency and reducing human errors. MetaTP leverages the Snakemake pipeline to achieve fully automated processing from data collection to differential expression analysis. Through automated management and efficient task scheduling, we successfully increased the efficiency of metatranscriptomic data analysis, providing researchers with a flexible and effective analysis tool. The innovation of metaTP lies in its high integration and automation in steps such as data collection, quality control, non-coding RNA removal, transcript expression quantification, and differential gene expression analysis, making the complex analysis workflow more convenient and user-friendly.

Results

Tool feature comparison

To assess the applicability of metaTP in metatranscriptomic analysis, we conducted a comparative evaluation with commonly utilized analysis pipelines, including HUMAnN [15], SAMSA2 [22], MetaTrans [23] and FMAP [24], as shown in Table 1. The comparison encompassed several key aspects: the analysis process, computational performance, functional coverage, and overall applicability.

MetaTP offers a comprehensive suite of integrated quality control and data preprocessing tools, such as FastQC and Trimmomatic, which effectively ensure data integrity. In contrast, tools like HUMAnN3 require users to manually manage these steps.

Table 1 Tool feature comparison

Comparative dimensions	metaTP	HUMAnN3	SAMSA2	MetaTrans	FMAP
Quality control	√FastQC, Trimmomatic	×(Need to be handled by the user)	√PEAR, Trimmomatic	√FastQC	√BMTagger
rRNA removal	√Bowtie2	×	√SortMeRNA	√SortMeRNA	√SortMeRNA
Transcript assembly	√MEGAHIT	×	×	×	×
Protein coding region prediction	√TransDecoder	×	×	√FragGeneScan	×
Transcript expression quantification	√Salmon	×	×	×	×
Differential gene expression analysis	√t-test	×	√DESeq2	√DESeq2	√Kruskal–Wallis
Functional annotation	√eggNOG, KEGG, GO	√rxn, GO, KO, level4ec, pfam, eggNOG	√DIAMOND	√eggNOG, MetaHIT, M5nr, SOAP2	√ODB3
Functional enrichment analysis	√clusterProfiler	√KEGG, MetaCyc	×	×	√KEGG, UniProt
Co-expression network analysis	√ggClusterNet	×	×	×	×
Computational efficiency	√Snakemake	×	×	×	×
Reproducibility	√Snakemake	×	×	×	×
Final analysis results	Gene expression + enrichment analysis + co-expression network	Functional annotation	Functional annotation	Gene expression + classification analysis	Differential abundance analysis + enrichment analysis

Furthermore, metaTP demonstrates notable advantages in areas such as transcript assembly, protein-coding region prediction, and transcript expression quantification. It integrates MEGAHIT for transcript assembly, TransDecoder for protein-coding region prediction, and supports the use of Salmon for expression quantification—features not typically found in other pipelines.

In terms of functional annotation, metaTP supports a range of databases, including eggNOG, KEGG, and GO, facilitating detailed functional annotations. Additionally, it incorporates functional enrichment analysis via clusterProfiler, which further enhances the depth of the analysis. Although other tools like SAMSA2 and FMAP also offer functional annotation, their capabilities for functional enrichment analysis are more limited. Moreover, metaTP integrates ggClusterNet for co-expression network analysis, enabling in-depth exploration of gene interactions, a feature that is often absent in other pipelines.

From a computational perspective, metaTP employs automated workflows based on Snakemake, supports parallel computing, and offers high computational efficiency and reproducibility, making it well-suited for processing large-scale datasets. In contrast, its computational efficiency and comprehensive workflow management provide greater flexibility and efficiency in handling complex analyses.

In summary, metaTP exhibits robust functionality across multiple facets of metatranscriptome analysis, particularly in data processing, functional annotation, expression quantification, and analytical efficiency. Its clear advantages make it particularly well-suited for research requiring multi-dimensional analysis and high-performance computing.

Case study

To evaluate the effectiveness of the metaTP in analyzing meta-transcriptome, eight bulk and rhizosphere metatranscriptomic samples from the published study [25] were employed for subsequent analysis. Exemplary output of metaTP for dimensionality reduction analysis (Fig. 2a), Venn analysis (Fig. 2b) and differential gene expression analysis (Fig. 2c) based on the soil microbiome dataset is summarized in Fig. 2. The PCoA scores demonstrated a significant separation between the rhizosphere samples and the different doses of bulk samples, indicating notable transcriptional differences in various soil environments. The Venn diagram provided a visual representation of the shared and

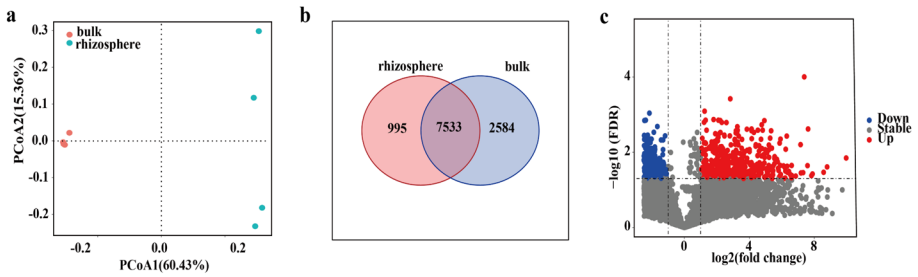

Fig. 2 Dimensionality reduction analysis (**a**), Venn analysis (**b**) and differential gene expression analysis (**c**) from metaTP. **a** A Principal Coordinate Analysis (PCoA) ordination is shown based on the Bray–Curtis distance matrix of gene expression profiles; **b** displays a Venn diagram illustrating the shared and unique genes; **c** Volcano plot diagram analysis of differentially expressed genes (DEGs)

unique genes between the bulk and rhizosphere soil. Figure 2b shows 995 unique genes for the rhizosphere soil and 7,533 shared genes. Additionally, the volcano plot depicted the differential expression of genes between the groups, with significantly upregulated or downregulated genes determined by Log 2 FC and log 10 FDR.

Functional profile analysis

The metaTP workflow provides an integrated analysis framework for the functional classification and enrichment analysis of differentially expressed genes (DEGs). By constructing sample-based local annotation packages, metaTP enhances the functional annotation capability for specific metatranscriptomic samples, improving target specificity compared to generic annotation databases and reducing interference from non-target information. After annotating genes with the eggNOG database using eggNOG-mapper, the metaTP generates annotation packages for the meta-transcriptomic samples. The annotation package is created locally using the 'makeOrgPackage' function from the R package 'AnnotationForge'. The generated annotation package contains gene information along with corresponding functional classifications, GO terms, and KEGG pathways, which play a crucial role in functional enrichment analysis of gene sets obtained from differential expression analysis. Gene Ontology (GO) and KEGG ontology (KO) enrichment was then performed using the clusterProfiler package in R. The GO functional analysis includes biological process (BP), cellular component (CC), and molecular function (MF). In the up-regulated genes of the rhizosphere samples, we found that the most enriched terms were cellular process (GO: biological process), metabolic process (GO: biological process), cellular anatomical entity (GO: cellular component), binding (GO: molecular function), and structural molecule activity (GO: molecular function) (Fig. 3a). These results reveal the functional characteristics of specific genes in rhizosphere samples. With the help of local annotation packages, metaTP enables precise mapping of DEGs to KEGG pathways, thereby providing a deeper understanding of gene functions. The identified up-regulated genes were classified into 14 function classes based on the Clusters of Orthologous Groups of proteins (COG) function classification. These classes include replication/recombination/repair (L), transcription (K), and translation/ribosomal structure/biogenesis (J), with the majority of genes falling into these groups, except for unknown function genes (Fig. 3b). Although there were differences in the relative abundance of genes annotated to the component of the MICOS complex and structural molecule activity between bulk and rhizosphere samples, no significantly enriched GO terms were observed in the enrichment analysis of up-regulated DEGs (Fig. 3c, d). Although the annotation package design of metaTP depends on the coverage of the eggNOG database, its flexibility and specificity for particular samples compensate for the limitations of generic tools to some extent, providing efficient support for the classification and annotation of functionally unknown genes.

Gene co-expression network

The metaTP pipeline involved converting the correlation matrix of all possible pairs of ASVs into an adjacency matrix. A cutoff of correlation coefficients was set at 0.6, with a significance threshold of $p < 0.001$ using the random matrix theory (RMT)-based method. The network was then constructed and visualized using ggClusterNet (Fig. 4a,

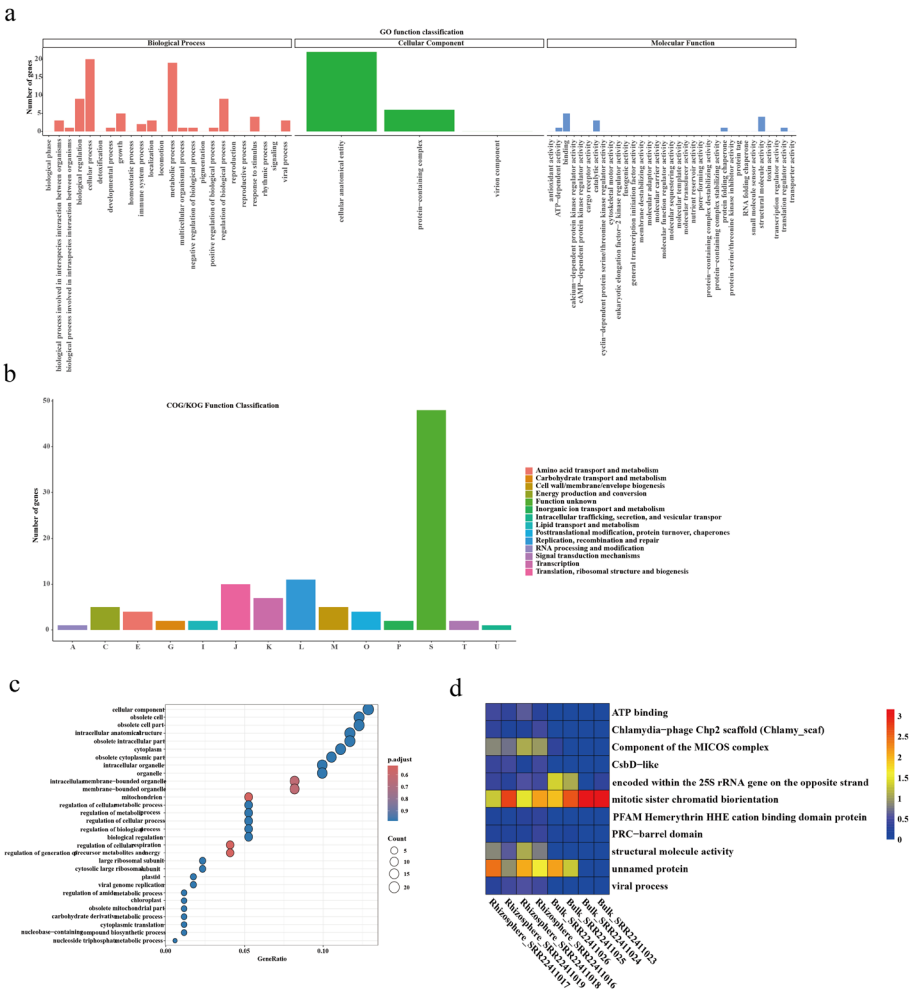

Fig. 3 Functional annotation and enrichment analysis of differentially up-regulated genes: **a** Gene Ontology (GO) functional enrichment analysis; **b** Clusters of orthologous groups of proteins (COG) function classification; **c** GO enrichment analysis for up-regulated DEGs; **d** Heatmap showing functional gene relative abundance between bulk and rhizosphere samples

b). In the network topologies, each node represents a gene. Degree represents the number of interaction partners for a node in a given network. The path length quantifies the level of integration in the network. Betweenness centrality is a measure of the number of shortest paths that go through a given node. Closeness centrality, on the other hand, measures the importance of a node and refers to how easily it can be reached from other nodes. In our case study, there was a high degree of average degree and density of the bulk network with respect to rhizosphere (Fig. 4e, l).

Conclusion

The study of microbiomes using meta-transcriptome sequencing enables the analysis of the activity of microbial communities that may have important roles in their environments. Over the past decade, omics technology has provided a theoretical foundation for understanding the distribution patterns and functional mechanisms of microorganisms under various conditions. Integrating different bioinformatic tools for

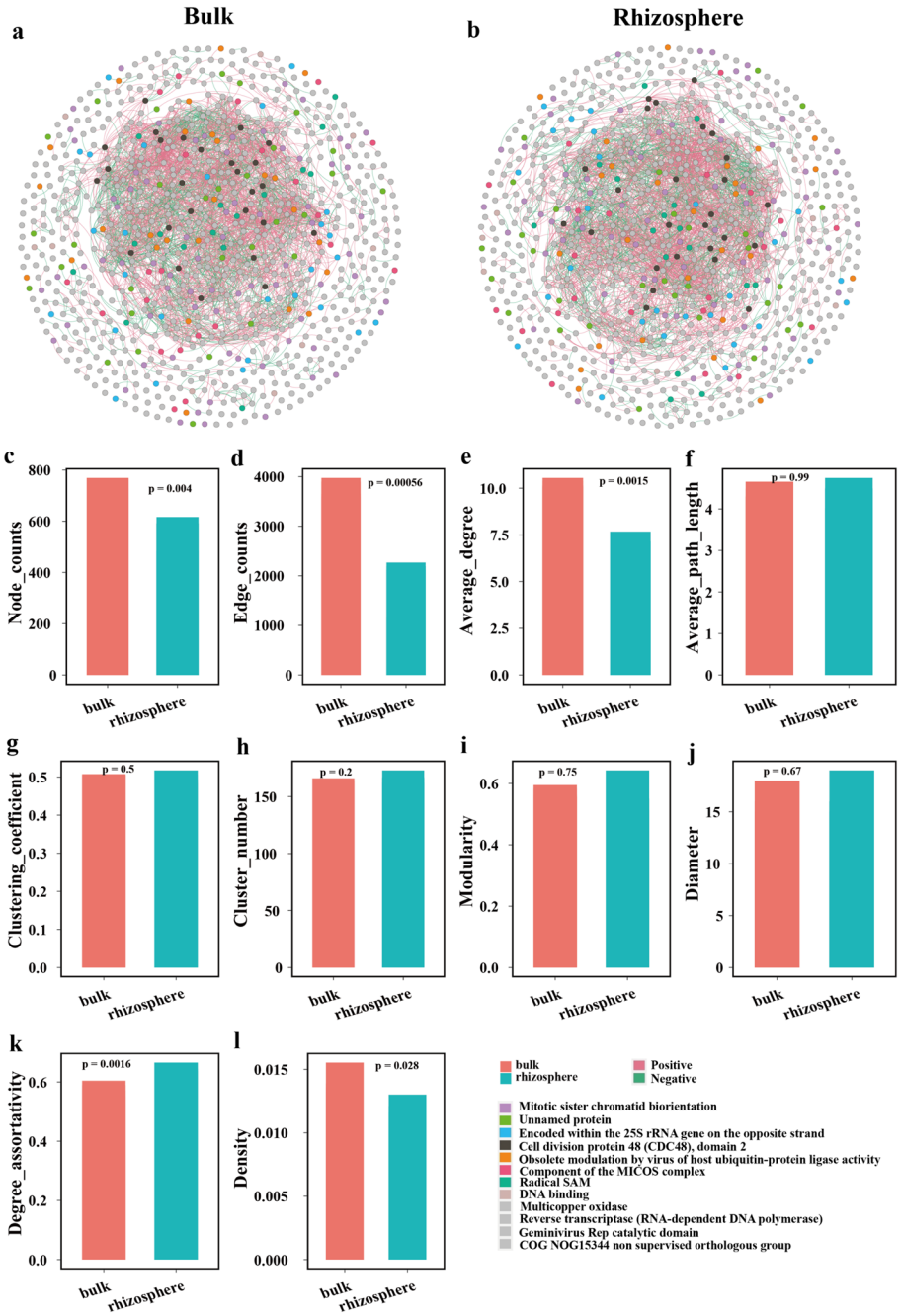

Fig. 4 Gene co-expression network analysis. Gene expression network of bulk (**a**) and rhizosphere (**b**) samples. **c**–**l** are the differences of network topology properties between bulk and rhizosphere samples

this high-throughput data has become the burden for biologists. The data processing pipeline has strong potential to improve the reproducibility in meta-analysis studies. In our study, metaTP provides an analytical environment with reproducible workflows that efficiently process raw data into a gene expression matrix with reference-independent quantification methods. MetaTP also provides downstream analysis and visualization methods including functional enrichment and gene co-expression network analysis,

which contain network topology calculations. Our hope is that this tool will serve in the future as a valuable resource for researchers.

Author contributions
LH, YW and QZ designed the study and prepared the manuscript; LH performed computational analyses and developed the pipeline. All authors read and approved the final manuscript.

Funding
Our work was supported by the National Natural Science Foundation of China (Nos. 62102269, 62373080).

Availability of data and codes materials
All test data and codes are shared on GitHub (https://github.com/nanbei45/metaTP).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 10 December 2023 Accepted: 8 April 2025
Published online: 26 April 2025

References

1. Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, Chen X, Cocolin L, Eversole K, Corral GH, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome. 2020;8(1):103.
2. Shi Y, Tyson GW, DeLong EF. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature. 2009;459(7244):266–9. https://doi.org/10.1038/nature08055.
3. France MT, Fu L, Rutt L, et al. Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data. Genome Biol. 2022;23(1):66. https://doi.org/10.1186/s13059-022-02635-9.
4. Booijink CCGM, Boekhorst J, Zoetendal EG, et al. Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed. Appl Environ Microbiol. 2010;76(16):5533–40.
5. Tveit AT, Urich T, Svenning MM. Metatranscriptomic analysis of arctic peat soil microbiota. Appl Environ Microbiol. 2014;80(18):5761–72.
6. Wei Z, Gu Y, Friman VP, et al. Initial soil microbiome composition and functioning predetermine future plant health. Sci Adv. 2019;5(9):eaaw0759.
7. Frank DN, Robertson CE, Hamm CM, et al. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. Inflamm Bowel Dis. 2011;17(1):179–84.
8. Gopalakrishnan V, Spencer CN, Nezi L, et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. Science. 2018;359(6371):97–103.
9. Lo C-C, Chain PS. Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinform. 2014;15(1):1–8.
10. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292–4.
11. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods. 2021;18(10):1161–8.
12. Ni Y, Li J, Panagiotou G. COMAN: a web server for comprehensive metatranscriptomics analysis. BMC Genom. 2016;17(1):622. https://doi.org/10.1186/s12864-016-2964-z.
13. Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform. 2008;9:386. https://doi.org/10.1186/1471-2105-9-386.
14. Narayanasamy S, Jarosz Y, Muller EEL, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol. 2016;17:1–21.
15. Franzosa EA, McIver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15(11):962–8.
16. Leinonen R, Sugawara H, Shumway M. Collaboration INSD: the sequence read archive. Nucleic Acids Res. 2010;39(suppl_1):D19–21.
17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
18. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
19. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–6.

20. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.
21. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38(12):5825–9.
22. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: a standalone metatranscriptome analysis pipeline. BMC Bioinform. 2018;19(1):175. https://doi.org/10.1186/s12859-018-2189-z.
23. Martinez X, Pozuelo M, Pascal V, et al. MetaTrans: an open-source pipeline for metatranscriptomics. Sci Rep. 2016;6:26447. https://doi.org/10.1038/srep26447.
24. Kim J, Kim MS, Koh AY, Xie Y, Zhan X. FMAP: functional mapping and analysis pipeline for metagenomics and metatranscriptomics studies. BMC Bioinform. 2016;17(1):420. https://doi.org/10.1186/s12859-016-1278-0.
25. Mendes LW, Raaijmakers JM, de Hollander M, Sepo E, Gómez Expósito R, Chiorato AF, Mendes R, Tsai SM, Carrión VJ. Impact of the fungal pathogen *Fusarium oxysporum* on the taxonomic and functional diversity of the common bean root microbiome. Environ Microb. 2023;18(1):68.

Publisher's Note