Differential gene expression in metatranscriptomics

Mariana Oliveira[1], João Carlos Sequeira[2] and Andreia Ferreira Salvador[2]

[1]School of Engineering, Minho University, Campus de Azurém, 4800-019,
Guimarães, Portugal
[2]Centre of Biological Engineering, Minho University, Campus de Gualtar, 4710 - 057, Braga, Portugal, direcao@ceb.uminho.pt
https://www.ceb.uminho.pt/

Abstract. This project involved a thorough review and analysis of statistical approaches for differential gene expression analysis in metatranscriptomic data. The primary focus was on exploring the ZIBSeq package, utilizing the LB test, and comparing it with the DESeq2 script within the MOSCA framework. The analysis encompassed an examination of the ZIBSeq package functionalities, investigation into the data origin from a previous study, and the execution of differential gene expression analysis on authentic MT data sourced from the Integrative Human Microbiome Project. The comparison between ZIBSeq and DESeq2 underscored the critical importance of selecting appropriate tools for managing zero counts in MT data. The study emphasized the necessity for optimizing analytical workflows and selecting suitable statistical approaches. By strategically selecting and integrating these tools, researchers can navigate the complexities of MT data, improve the quality of analyses, and contribute to a more comprehensive understanding of microbial community dynamics.

Keywords: differential gene expression, metatranscriptomics, MOSCA, DESeq2, logistic-beta test

1 Introduction to Metatranscriptomics and Statistical Challenges

The functional dynamics of a microbial community can be evaluated through the presence of expressed genes, or transcriptomes, which are captured by metatranscriptomics (MT). This essential technique allows for the study and analysis of the functional profile, physiology, and structure of unknown microbial communities by capturing real-time mRNA expression in selected samples, thus allowing the disclosure of particulars about the population which are transcriptionally active [1]. This technique has far-reaching applications across various academic disciplines, with medicine standing out as a major beneficiary. The analyses of disease-associated communities have become one of the most important fields within MT allowing for an increased knowledge on the importance of the interactions between microbes and host in human health [2].

Even though the MT sequencing data has been growing in importance and accessibility, there has been limited progress in the development of specialized statistical analysis techniques tailored to this field [3]. Most of the approaches used for analyzing MT data rely on readily available tools that have undergone adjustments to suit the characteristics of gene analysis within microbial community context like the adaptation of analytic pipelines that were originally designed for other high-throughput sequencing technologies, such as single-organism RNA-seq, or that were initially developed for analyzing 16S ribosomal RNA [3, 4].

The study of MT must go through multiple stages including: 1) aligning or assembling transcripts; 2) annotating transcripts with functional and/or taxonomic details; 3) normalizing the data; and 4) conducting differential expression analysis [4]. The latter allows for the comparison of gene expression patterns to identify, within a microbial community, its active elements and evaluate microbial changes [2, 5]. Differential expression analysis allows for the comparison of expression levels of taxa or microbial genes between samples [3].

When typical statistical inference methods are used on MT data, false positives or negative rates may be exacerbated due to the high dimensionality, sparsity and mean-variance nature of the data [3]. Additionally, the presence of zero inflation presents challenges in the MT data analysis, due to the extensive number of features, their dynamic range and the detection limits imposed by sequencing depth [6]. Therefore, it’s important to account for the characteristics of microbiome count data during statistical analysis to fix eventual irreproducible associations [4].

2 Differential Gene Expression Analysis: Evolution and Approaches

The methods initially employed to conduct differential gene expression (DGE) relied on calculating fold change, aiming to measure the magnitude of gene expression alterations between experimental conditions, such as treatment and control groups. However, fold change analysis has inherent limitations, including susceptibility to noise in estimates, reliance on experimental design, and challenges in interpreting small changes. Subsequently, statistical methods emerged as a preferred approach to assess quantitative differences between experiments. This shift was driven by the need to enhance the robustness and biological relevance of fold change analyses. By integrating statistical frameworks, researchers aimed to address the limitations of fold change analysis and provide more accurate assessments of gene expression changes across experimental conditions [7].

2.1 Parametric vs Non-Parametric Approaches

All the methods that have been used and developed for the studying of MT data can be separated into two major groups - parametric and non-parametric. Parametric ones are used in tools such as metagenomeSeq and MaAsLin which are some examples of community-specific tools; apart from these, some methods used for differential expression analysis data have also been adopted by other tools – such as edgeR, DESeq2 and voom(+limma) - which are not community-specific tools. Non-parametric ones are included in SAMseq, NOIseq, LEfSe and ANCOM [4, 8]. Parametric methods encapsulate data characteristics within predefined parameters, allowing for predictions of unknown data based on the model and its parameters. Typically, these methods assume a specific distribution, such as Poisson or negative binomial (NB), after normalization. In contrast, non-parametric methods offer greater flexibility by not imposing rigid distributional assumptions. They capture more nuanced data distributions, recognizing that data characteristics may not be fully captured by a finite set of parameters. This flexibility allows non-parametric models to adapt to varying data volumes, potentially offering deeper insights into gene expression dynamics [9].

In the following subsections, a comprehensive analysis and review of various commonly applied tools for DGE will be provided. This review will outline the advantages and limitations of each one. For further statistical information and detailed comparison of these tools, please refer to Table 1 in the appendix.

Parametric Models: Poisson and Negative Binomial. The Poisson distribution and the NB distribution are the most used models for parametric DGE. The Poisson distribution, characterized by its simplicity and single parameter, imposes a constraint where the variance equals the mean of the modeled variable. In contrast, the NB distribution, featuring two parameters representing mean and dispersion, offers flexibility by accommodating a wider range of mean-variance relationships. The NB model is typically employed to address overdispersion in the data [10].

A commonly used tool that employs the Poisson distribution is **DEGseq** and another method that also makes use of this distribution is the Two-stage Poisson Model (**TSPM**). **DEGseq** assumes a binomial or Poisson distribution, making it efficient for small-scale studies but limited to datasets without overdispersion [11]. Despite this constraint, DEGseq provides an easy-to-use solution for identifying differentially expressed genes [12]. In the meantime, **TSPM** has shown flexibility for complex studies and sensitivity to sample size and data with overdispersion [10].

Meanwhile, the NB distribution is present in the following tools: **baySeq**, **Cuffdiff2**, **DESeq**, **DESeq2**, **EBSeq**, **edgeR**, **NBPseq** and **vst(+limma)**. **BaySeq**, **DESeq2** and **EBSeq** have demonstrated relatively low/conservative false discovery rates (FDR), whereas **Cuffdiff2**, **edgeR** and **NBSeq** exhibits a notably higher FDR [8, 10]. **Cuffdiff2** can only be applied to two-group differential expression analyses, while **vst(+limma)** requires at least 3 sample sizes to detect DGE [10, 13]. Both **edgeR** and **Cuffdiff2** have shown a high sensitivity and **bayseq** has shown a high variability within its results [8, 11, 14]. Regarding computational demands, **bayseq** is slow, **vst(+limma)** is fast and **Cuffdiff2** shows computational demands [10, 13, 15]. Additionally, both **vst(+limma)** and **DESeq2** are efficient for large sample sizes and, in the opposite direction, **EBSeq** is efficient for small sample sizes [8, 10, 16].

Parametric Models: Alternative Approaches. The **voom(+limma)** is an alternative package to vst(+limma), as it's coupled with the voom transformation. This tool is computationally efficient and robust but necessitates a minimum of 3 samples per condition [10, 17]. **ShrinkSeq** allows users to choose distributions, evaluating posterior probabilities for inference. It effectively shrinks dispersion parameters, ensuring a high true positive rate, yet it's computationally intensive [10, 18].

Non-Parametric Models. Some non-parametric approaches have also been commonly used for DGE, such as the **NOIseq**, **NOIseqBio** and **SAMseq**. **NOIseq** is effective in avoiding false positives, making it suitable for studies with varying numbers of replicates. However, its computational time requirement can be significantly influenced by the size of the sample [10, 11]. **NOISeqBIO** enhances gene-specific handling of biological variability, effectively controlling high FDR in experiments with biological replicates [19]. **SAMseq** performs effectively in controlling FDR while maintaining acceptable sensitivity but requires a minimum of 4-5 samples per condition [11, 17, 20].

2.2 Specific Tools for Microbiome Differential Gene Expression Analysis

Parametric Models. Some parametric models are used in tools commonly used for DGE that were specifically developed for microbiome data analysis. These include **ALDEx2**, **MaAsLin, metagenomeSeq** and **ZIBSeq** [3, 4]. **ALDEx2** employs a model with various hypothesis testing approaches, optimizing analysis for datasets with three or more replicates [21, 22]. This approach also enables the calculation of expected false FDR and adjusted *p*-values per transcript [23]. **MaAsLin** ranks factors based on their contribution to microbiota differences, it's especially useful in scenarios with multiple factors and can also perform univariate association [4, 24]. Meanwhile, the normalization employed (Total Sum-Scaling (TSS)) may introduce bias [25]. As for **metagenomeSeq**, it has demonstrated superior performance compared to tools like LEfSe, making it valuable for microbiome analysis [26, 27]. It has also shown a high sensitivity in most scenarios tested by Cho et al. (2023), although it doesn't account for batch effects [3]. In metatranscriptomics, batch effects can be particularly problematic due to the complexity of the data and the high variability in gene expression levels. For instance, the high cost of high-throughput profiling experiments or the difficulty in collecting a good number of samples can lead to small sample sizes, which can exacerbate batch effects [28]. Additionally, the use of different sequencing technologies, library preparation methods, or experimental conditions can introduce batch effects that need to be addressed [29].

Similarly, **ZIBSeq** exhibits high sensitivity in most scenarios, but caution is necessary when using it with a high proportion of zeros in small sample sizes, as it may result in an inflated type I error [3]. ZIBSeq also offers advantages such as efficient computation time and the ability to handle excess zeros, making it a valuable tool for

DE analysis in MT data [3]. According to Cho et al., the Logistic-Beta (LB) test employed by **ZIBSeq** is one of the highest-performing statistical methods, surpassing ALDEx2, MaAsLin, and metagenomeSeq [3].

Non-Parametric Models. Several non-parametric alternatives have also been developed specifically for microbiome data analysis, such as **ANCOM** and **LEfSe** [4]. **ANCOM** handles large datasets efficiently and performs well in controlling FDR, if the sample size isn't too small [27, 30]. In the meantime, it lacks *p*-values for individual taxa, standard errors or confidence intervals for differential abundance [27]. **LEfSe** has a relative high mean pairwise concordance and is effective in controlling false positives but may have reduced sensitivity and face limitations with complex datasets [31, 32].

Tools Performance. Comprehensive analysis and evaluation of the tools employed in MT DGE has been conducted, yet there has been scant review regarding the tools performance at the gene or gene-family level [3]. Instead, they commonly evaluate differential abundance or differential expression approaches at the species or at the taxon-level [27, 33, 34].

Cho et al. (2023) recently conducted a comparative evaluation of statistical analysis approaches used in microbial analyses in MT [3]. They assessed recently developed models designed to address zero-inflated over-dispersed counts or compositional data. These models are present in tools such as Model-based Analysis of Single-cell Transcriptomics (MAST) and ZIBSeq. The evaluation focused on tool/model performance at the gene or gene-family level in MT, aligning with recent trends in microbiome differential analysis prioritizing flexibility to handle high proportions of zeros in microbiome data.

The study identified the Log-normal test and the LB test, used by ZIBSeq, as the top-performing statistical methods. Overall, the Log-normal test effectively controls type I error and FDR. LB demonstrated superior sensitivity across various scenarios, although caution is advised due to potential inflated type I error with high zero proportions in small sample size. Notably, larger sample sizes and higher proportions of signal genes result in increased sensitivity for both methods, and they can maintain type I error and FDR at or below the nominal level of 5%. Moreover, considering computational efficiency, the Log-normal test and LB stand out as top-performing approaches.

In the same study, DESeq2 showed a very low type I error when used to model the zero-inflated data. However, it often has lower than 5% sensitivity when the data are sparse. This can be explained by DESeq2's inability to model zero-inflation.

3 Advancing Automated Analysis in Metatranscriptomics: The MOSCA Framework

Novel developments in the analysis of MT data have enabled the introduction of a new framework, Meta-Omics Software for Community Analysis (MOSCA). This pipeline was created for automated and integrated analysis of metagenomics (MG) and MT data performing preprocessing of raw files, assembly, annotation, DGE and comparison of multiple samples [35]. It performs DGE with Bowtie2, aligning the reads to the scaffolds obtained after the assembly [13]. For MG reads, normalization involves dividing read counts by scaffold length, then further normalizing using edgeR with the choice of Trimmed Mean of M-values (TMM) or Relative Log Expression (RLE) methods [36]. MT reads are divided by gene length and normalized similarly.

The DGE conducted by MOSCA makes use of the DESeq2 R package with MT read counts as input [35]. DESeq2 provides *p*-values indicating whether gene expression differs significantly between samples.

4 Objective and motivation

The main objective is to improve the MOSCA framework by refining its approach to DGE analysis for MT data through the implementation of a cut-to-edge solution, to specifically tackle the problem of zero-inflation modeling.

5 Methodology

5.1 Review of tools used in metatranscriptomic DGE

A comprehensive review of various approaches used in the field of DGE analysis was conducted (section 2), focusing on assessing the statistical aspects, the strengths and limitations of each methodology within the field of meta-omics analysis. This review considered parameters such as normalization, distribution, hypothesis testing, and dispersion estimation.

5.2 Testing and exploring the ZIBSeq and LB test: a novel approach in MOSCA framework

The next step involved exploring ZIBSeq, which conducts the LB test, by analyzing and exploring its package and understanding its functions. This exploration involved reading the package documentation and using its functions in RStudio.

Furthermore, we examined the origin and characteristics of the data used by Cho et al. (2023) [3]. This involved researching the background and origin of the data by consulting the original article from which it was derived. Additional analysis of this data involved exploration of their datasets using RStudio.

The tools used to generate the data used by Cho et al. (2023) were also explored [3]. This involved getting to know the workflow that was conducted to generate the data and gaining insight on the tools that were used in this workflow.

Subsequently, we proceeded with an in-depth analysis of the LB test script itself.

5.3 Comparison of Logistic-Beta test and DESeq2 with real data

Finally, we conducted DGE analysis on real MT data using both the DESeq2 script from MOSCA and the adapted LB test script. This dual approach allowed us to further compare the results obtained, highlighting key differences in their performance and outcomes.

The data used consisted of enzyme commission numbers (ECS) from the MT data of the Integrative Human Microbiome Project [37]. This data was used to study the association between gut microbiome and inflammatory bowel diseases (IBD). It includes metatranscriptomes of fecal samples from 132 subjects for a 1-year period, with repeated measurements of the same participants over time. It also included associated metadata, every sample has associated data regarding, for example, gender, age and diagnosis – ulcerative colitis (UC), Chron’s Disease (CD) and non-IBD.

For the DGE analysis conducted, we categorized subjects into "healthy" (non-IBD) and "non-healthy" groups, where the latter included patients with CD and UC. This classification reflects the clinical diversity among study participants.

6 Results and Discussion

6.1 Review of tools used in metatranscriptomic DGE

Based on the outcomes of the comprehensive review that was conducted, it was determined that replacing DESeq2 with ZIBSeq in the MOSCA framework would be an interesting change to study and evaluate (Figure 1).

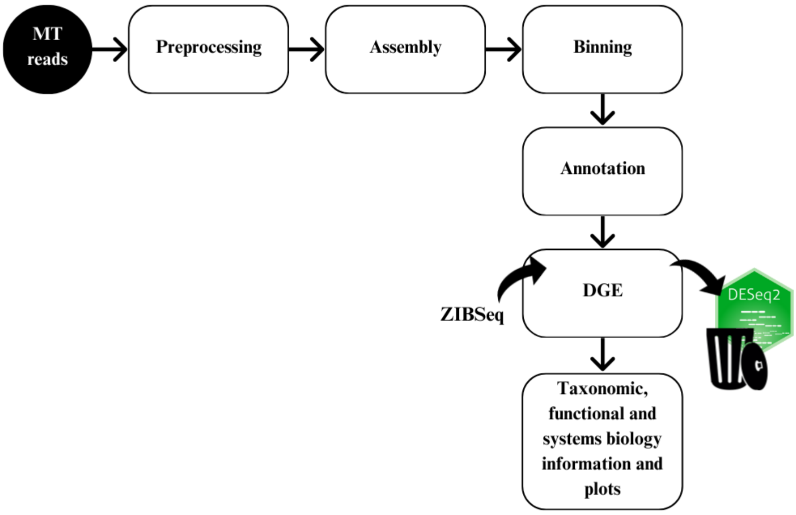

MT reads
Preprocessing
Assembly
Binning
Annotation
DGE
ZIBSeq
DESeq2
Taxonomic, functional and systems biology information and plots

Fig. 1. Illustration of MOSCA scheme, with proposed modifications indicated by large arrows.

ZIBSeq, designed for microbiome data, effectively addresses the compositional nature of the data, enabling it to handle zero counts. This stands in contrast to DESeq2, initially designed for bulk RNAseq analysis, which does not account for zero counts [3, 38]. This aspect is particularly pertinent for conducting DGE on MT data because, as previously mentioned, the prevalence of zeros in MT data complicates the selection of an appropriate tool - one that effectively manages the abundance of zero counts [3, 4]. This change is aimed at enhancing the framework's workflow and addressing the considerations identified during the review process ultimately optimizing its performance for analyzing meta-omics, in particular MT data.

6.2 Testing and exploring the ZIBSeq and LB test: a novel approach in MOSCA framework

ZIBSeq package. Upon inspection of the ZIBSeq, we observed that the package had been recently removed from the CRAN repository due to unresolved issues despite reminders. Nevertheless, the archived version 1.2 was still available and could be used for our analysis.

Our investigation revealed that the ZIBSeq package does feature screening, data normalization, zero-inflated beta regression and multiple hypothesis testing [38]. Furthermore, the ZIBSeq code uses the gamlss package. This package implements the zero-inflated beta regression model by numerically finding the maximum likelihood estimates of its parameters [3, 39].

ZIBSeq comes with two functions and features a real metagenomic dataset from Zupancic et al. (2012) [40]. This dataset comprises 310 cases and 248 variables. Among the 248 variables, 240 represent taxa at the genus level, and 8 correspond to clinical phenotypes. It provides a simulated scenario for exploring relationships between health-related variables (such as BMI, cholesterol, and glucose) and the counts or measurements of microbial species. This is valuable for testing and demonstrating statistical methods for analyzing complex metagenomic data.

Our analysis of this dataset with ZIBSeq initially involved the comparison of categorical variables without applying a square root transformation, identifying only *Faecalibacterium*. However, when the square root transformation was applied, both *Faecalibacterium* and *Ruminococcus* were identified, aligning our results with those reported by Peng et al. (2016). In contrast, when analyzing ordinal variables, our results did not entirely match those of Peng et al. (2016). Specifically, their analysis identified an additional taxon, *Collinsella*, which we were unable to detect. This discrepancy in results compared to Peng et al. (2016) may arise from the unresolved issues affecting data analysis, which could have implications for the interpretation and outcomes of our study [38].

Logistic-Beta test. Cho et al. (2023) conducted a comprehensive evaluation using three datasets sourced from two studies focusing on the human microbiome. The first two datasets, ZOE-pilot and ZOE2.0, originated from a molecular epidemiological investigation targeting early childhood caries study [41, 42]. Both datasets are available online, but accessing them requires a formal request, which can be a time-consuming process. Due to these procedural constraints, we were unable to use this data for further investigation into the script of the article. The third dataset originated from the same study we used to evaluate both approaches – DESeq2 and LB test -, the study that investigated the relationship between the gut microbial ecosystem and IBD. This third dataset is publicly available at the Inflammatory Bowel Disease Multi’omics Data Base (IBDMDB).

According to the authors, all three datasets’ raw metagenomic and metatranscriptomic sequencing reads were submitted to the bioBakery workflow, generating, for each sample a file of gene family abundances, pathway abundance and pathway coverage. Additionally, these files were regrouped into ECS and normalized to relative abundances (relab) (Figure 2) [43].

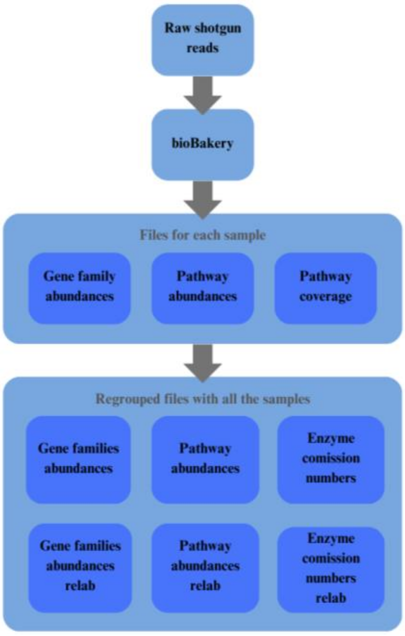


Fig. 2. Overview of data generation using the bioBakery workflow.

The files sourced from the IBDMDB, generated using the bioBakery workflow, presented challenges during our analysis. While attempting to reproduce results and comprehend the application of the LB test script, we encountered discrepancies in file names and descriptions. Furthermore, our exploration of the bioBakery workflow on GitHub, which was another data source, revealed that it only contained taxonomical and functional profiles for the MG data. The absence of MT data in this repository further complicated our efforts and required extensive time investment.

Afterwards, we explored the LB test function implemented by the authors. It begins by computing the proportion of counts for each sample from the provided data. It then employs the gamlss function to fit a model, making necessary adjustments during the process. Subsequently, the function extracts coefficients and calculates *p*-values for the effect of conditions. Finally, it returns an output matrix containing estimates and corresponding p-values derived from the model. Although the function successfully processed the taxonomical profiles data from the IBDMDB and conducted comparisons between the conditions "Sick vs Healthy" and "Antibiotic vs No Antibiotic," it encountered difficulties in generating the expected output for differentially expressed genes. Further investigation led to the discovery of the script designed specifically to obtain this crucial information.

The script to obtain the DGE begins by preparing gene expression data and associated metadata. It encodes gender information into binary values for analysis. Using a screening function, it identifies genes meeting specified expression criteria and selects a subset for further analysis. Each selected gene undergoes logistic regression modeling to examine its relationship with the chosen conditions. The script iteratively processes genes, collecting and organizing results for further interpretation and analysis.

6.3 Comparison of Logistic-Beta test and DESeq2 with real data

Using the LB test, we identified 615 significant ECS in the dataset. In contrast, applying DESeq2 with a fold change threshold of 1 (testing for the hypothesis that there was some difference between the samples) yielded 3017 significant ECS. Of these, only 158 ECS were common to both methods (Figure 3). When applying DESeq2 with a fold change threshold of 2 (testing for the hypothesis that the enzymes mapped to that EC had at least double the expression from one sample to the other), we identified 52 significant ECS in the dataset, of which only 10 matched the results from the LB test (Figure 4).

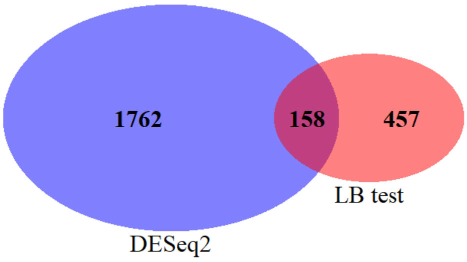


Fig. 3. Venn diagram comparison of significant ECS identified by LB Test and DESeq2 with fold change of 1.

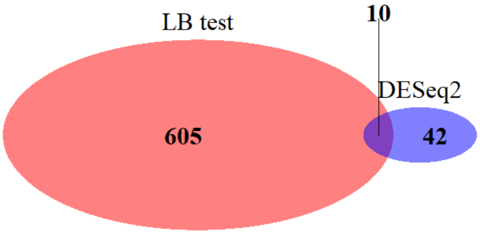


Fig. 4. Venn diagram comparison of significant ECS identified by LB Test and DESeq2 with fold change of 2.

This analysis highlights the different outcomes between LB test and DESeq2 approaches. While the LB test uses a logistic-beta distribution to model the gene expression data, DESeq2 employs a negative binomial distribution adjusted for over-dispersion in count data [3, 11]. These differences may influence the detection of significant ECS, perhaps leading to varying numbers of identified ECS in our project. Moreover, as mentioned by Cho et al. (2023), DESeq2 showed overall worse levels of sensitivity when compared to the other methods evaluated by the authors [3]. This can further explain why DESeq2, in our first analysis, estimated more significant ECS compared with LB test. Additionally, in our second analysis, when applying a fold change threshold of 2, DESeq2 exhibits a substantial reduction in the number of identified differentially expressed ECS, indicating a stricter criterion for significant expression changes.

In addition, the DESeq2 in MOSCA provides a table of normalized counts as an output and afterwards, the MOSCA produces graphical representations such as MA plots and principal component analysis (PCA) plots to visualize data distribution and variability across conditions (Figure 5). The LB test has not yet been adapted to provide this information. It primarily presents a list of significant genes without accompanying graphical explorations (Figure 6), which facilitates a quick identification and downstream functional analysis.

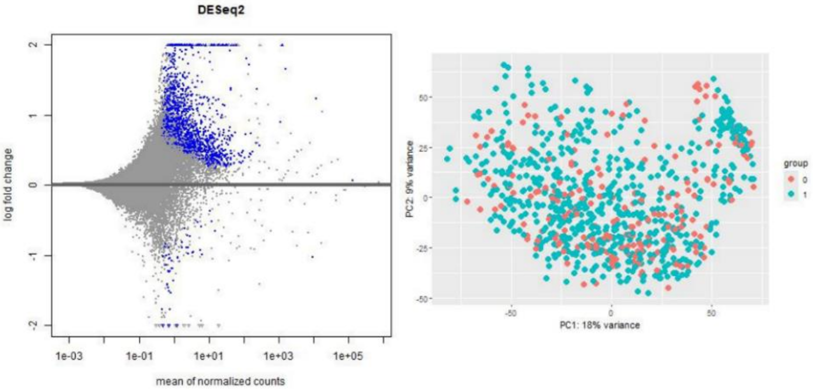
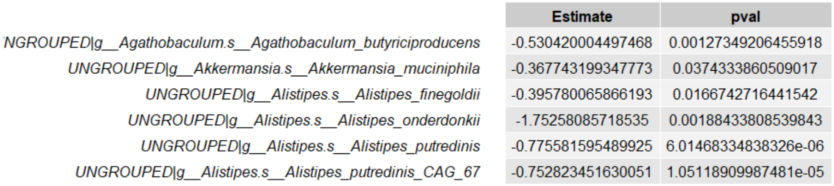


Fig. 5. MA plot and PCA generated by the workflow in MOSCA.



	Estimate	pval
NGROUPED|g__Agathobaculum.s__Agathobaculum_butyriciproducens	-0.530420004497468	0.00127349206455918
UNGROUPED|g__Akkermansia.s__Akkermansia_muciniphila	-0.367743199347773	0.0374333860509017
UNGROUPED|g__Alistipes.s__Alistipes_finegoldii	-0.395780065866193	0.0166742716441542
UNGROUPED|g__Alistipes.s__Alistipes_onderdonkii	-1.75258085718535	0.00188433808539843
UNGROUPED|g__Alistipes.s__Alistipes_putredinis	-0.775581595489925	6.01468334838326e-06
UNGROUPED|g__Alistipes.s__Alistipes_putredinis_CAG_67	-0.752823451630051	1.05118909987481e-05

Fig. 6. Subset of output generated by the LB test.

Our analysis had several limitations. The results were influenced by the different statistical approaches used, as the LB test and DESeq2 rely on distinct distribution assumptions and statistical models, which in turn affected the detection of significant ECS. Using simulated data with known expression profiles would be beneficial for gaining a deeper understanding of which approach performs better. This would enable us to better evaluate the sensitivity and specificity of each method under controlled conditions, providing clearer insights into their respective performances.

Given these considerations, it is premature to assert that one approach is superior to the other. Further analyses, including the use of simulated data, are necessary to better understand the strengths and weaknesses of each method.

Furthermore, future research could focus on integrating the strengths of both approaches. For instance, combining the LB test's straightforward output with DESeq2's visual tools could enhance the overall analysis process, providing both clarity in results and deeper insights into data variability and distribution.

7 Conclusion

The comprehensive review and analysis conducted in this project highlights the importance of selecting appropriate statistical approaches to the characteristics of MT data. Through the exploration of different approaches, valuable insights have been gained into their capabilities and performance in handling DGE analysis in MT.

Moreover, optimizing analytical workflows by integrating tools which effectively address the compositional nature of MT data and manage zero counts is crucial. This strategic approach not only enhances the accuracy and reliability of DGE analysis in MT but also advances our understanding of microbial community dynamics. By making use of suitable statistical approaches and tools, researchers can make more informed decisions in future MT studies, thereby advancing analytical methodologies and gaining deeper insights into complex biological systems.

8 Appendices

Table 1. Differential gene expression analysis methods.

Tool	Normalization	Distribution	Hypothesis test	Dispersion estimation	Ref.
ALDEx2	CLR	Dirichlet-multinomial	Wilcoxon Rank Sum/Welch's t-test, Kruskal-Wallis/correlation	Standard Bayesian techniques with Dirichlet prior	[3, 23]
ANCOM	ALR	Non-parametric	Mann-Whitney U	Not specified	[4]
baySeq	Scaling factors (quantile/TMM/total)	NB	Evaluating posterior probability for inference	Maximum likelihoo d	[10, 11, 15]
Cuffdiff2	Geometric/FPKM	Beta NB	*t*-test analogical method	Maximum likehood	[13, 20, 44]
DESeq	DESeq size factors	NB	Classical	Modeling mean-variance relationship using parametric or local regression	[8–10]
DESeq2	DESeq size factors	NB with GLM	Wald	Empirical Bayes	[8, 9, 11]
DEGseq	None, Loess, Median	Binomial or Poisson	MARS, FET, LRT	Sliding-window	[11, 12]

Tool	Normalization	Distribution	Hypothesis test	Dispersion estimation	Ref.
EBSeq	DESeq median/quatile	NB	Bayesian (Fisher's exact test)	Expectation-estimation algorithm	[8, 9, 16]
edgeR	TMM/Upperquartile /RLE/None (all scaling factors are set to be one)	Negative Binomial (Exact or GLM)	Classical	Maximum likelihood	[8, 9, 11]
LEfSe	TSS	Non-parametric	Kruskal-Wallis, Wilcoxon rank-sum, Linear Discriminant Analysis	Not specified	[4, 31, 45]
MaAsLin	TSS	Gaussian	Wald	Not specified	[4]
metagenomeSeq	CSS	Zero-inflated Gaussian	Moderated *t*	Mixture model with ZIG distribution	[4, 27]
NBPSeq	None (all scaling factors are set to be one)/DESeq like	NB	Classical	Modeling mean-variance relationship using parametric or local regression	[10, 17]
NOIseq	RPKM/TMM/Upperquartil e	Non-parametric	Ratio of fold change and absolute expression differences	Not specified	[9, 11, 20]
NOIseqBIO	RPKM/TMM/Upperquartil e	Non-parametric	Statistical method based on Z-statistic	Empirical Bayes	[19]
SAMseq	Based on the read count mean over the null features of data set	Non-parametric	Wilcoxon rank su	Permuation	[9, 17, 20]
ShrinkSeq	TMM/Not specified	User can select different distributions	Evaluating posterior probability for inference	Shrinkage of dispersion parameter	[10, 18]

Tool	**Normalization**	**Distribution**	**Hypothesis test**	**Dispersion estimation**	**Ref.**
TSPM	TMM/Not specified	Two-stage Poisson	Likehood ratio/Quasi-likehood approach	Approximated Chi-squared distribution or F-distribution	[10, 17, 46]
Voom(+ limma)	TMM	Gaussian/GLM	Moderated *t*/ moderated F	Modeling mean-variance relationship	[10, 17]
Vst(+ limma)	Division by size factors or normalization factors	NB	Moderated *t*/ moderated F	Modeling mean-variance relationship	[10, 17, 47]
ZIBseq	TSS	Zero-inflated beta	Wald/likehoo d	Maximum likehood	[48, 49]

References

1. Dubey, R.K., Tripathi, V., Prabha, R., Chaurasia, R., Singh, D.P., Rao, Ch.S., El-Keblawy, A., Abhilash, P.C.: Metatranscriptomics and Metaproteomics for Microbial Communities Profiling. Presented at the (2020)
2. Ojala, T., Kankuri, E., Kankainen, M.: Understanding human health through metatranscriptomics, (2023)
3. Cho, H., Qu, Y., Liu, C., Tang, B., Lyu, R., Lin, B.M., Roach, J., Azcarate-Peril, M.A., Aguiar Ribeiro, A., Love, M.I., Divaris, K., Wu, D.: Comprehensive evaluation of methods for differential expression analysis of metatranscriptomics data. Brief Bioinform. 24, (2023). https://doi.org/10.1093/bib/bbad279
4. Mallick, H., Ma, S., Franzosa, E.A., Vatanen, T., Morgan, X.C., Huttenhower, C.: Experimental design and quantitative analysis of microbial community multiomics, (2017)
5. McDermaid, A., Monier, B., Zhao, J., Liu, B., Ma, Q.: Interpretation of differential gene expression results of RNA-seq data: Review and integration, (2019)
6. Zhang, Y., Thompson, K.N., Huttenhower, C., Franzosa, E.A.: Statistical approaches for differential expression analysis in metatranscriptomics. Bioinformatics. 37, I34–I41 (2021). https://doi.org/10.1093/bioinformatics/btab327
7. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, (2014). https://doi.org/10.1186/s13059-014-0550-8
8. Li, D., Zand, M.S., Dye, T.D., Goniewicz, M.L., Rahman, I., Xie, Z.: An evaluation of RNA-seq differential analysis methods. PLoS One. 17, (2022). https://doi.org/10.1371/journal.pone.0264246
9. Costa-Silva, J., Domingues, D., Lopes, F.M.: RNA-Seq differential expression analysis: An extended review and a software tool, (2017)
10. Soneson, C., Delorenzi, M.: A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 14, (2013). https://doi.org/10.1186/1471-2105-14-91
11. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A.: A survey of best practices for RNA-seq data analysis, (2016)
12. Gao, D., Kim, J., Kim, H., Phang, T.L., Selby, H., Choon Tan, A., Tong, T.: A survey of statistical software for analysing RNA-seq data. (2010)
13. Frazee, A.C., Pertea, G., Jaffe, A.E., Langmead, B., Salzberg, S.L., Leek, J.T.: Ballgown bridges the gap between transcriptome assembly and expression analysis, (2015)
14. Alshehri, H., Alkharouf, N.: Compare and contrast of differential gene expression software packages of RNA-Seq. In: Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018. pp. 1374–1379. Institute of Electrical and Electronics Engineers Inc. (2018)

15. Seyednasrollah, F., Laiho, A., Elo, L.L.: Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 16, 59–70 (2013). https://doi.org/10.1093/bib/bbt086
16. Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M.G., Haag, J.D., Gould, M.N., Stewart, R.M., Kendziorski, C.: EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 29, 1035–1043 (2013). https://doi.org/10.1093/bioinformatics/btt087
17. Khadka, V.S., Wang, L., Lim, E., Khadka, V.S., Chen, J.J.: Comparison of False Positive in Tools for Differential Gene Expression Calling in RNA-Seq Analysis. (2016)
18. Huang, H.C., Niu, Y., Qin, L.X.: Differential expression analysis for RNA-Seq: An overview of statistical methods and computational software. Cancer Inform. 14, 57–67 (2015). https://doi.org/10.4137/CIN.S21631
19. Tarazona, S., Furió-Tarí, P., Turrà, D., Di Pietro, A., Nueda, M.J., Ferrer, A., Conesa, A.: Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 43, (2015). https://doi.org/10.1093/nar/gkv711
20. Li, D.: Computational Biology. Presented at the November 1 (2019)
21. Gloor, G., Fernandes, A., Macklaim, J., Albert, A., Links, M., Quinn, T., Wu, J., Wong, R., Lieng, B., Nixon, M.: Bioconductor, https://bioconductor.org/packages/release/bioc/html/ALDEx2.html
22. Fernandes, A.D., Macklaim, J.M., Linn, T.G., Reid, G., Gloor, G.B.: ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLoS One. 8, (2013). https://doi.org/10.1371/journal.pone.0067019
23. Quinn, T.P., Crowley, T.M., Richardson, M.F.: Benchmarking differential expression analysis tools for RNA-Seq: Normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics. 19, (2018). https://doi.org/10.1186/s12859-018-2261-8
24. Odintsova, V., Tyakht, A., Alexeev, D.: Guidelines to Statistical Analysis of Microbial Composition Data Inferred from Metagenomic Sequencing. Curr Issues Mol Biol. 24, 17–36 (2017). https://doi.org/10.21775/cimb.024.017
25. Ramadhan, M.R.: Comparison of Metagenomic Tools in Gut Microbiome Analysis of COVID-19 Patients. (2022)
26. Paulson, J.N., Colin Stine, O., Bravo, H.C., Pop, M.: Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 10, 1200–1202 (2013). https://doi.org/10.1038/nmeth.2658
27. Lin, H., Peddada, S. Das: Analysis of compositions of microbiomes with bias correction. Nat Commun. 11, (2020). https://doi.org/10.1038/s41467-020-17041-7
28. Li, T., Zhang, Y., Patil, P., Johnson, W.E.: Overcoming the impacts of two-step batch effect correction on gene expression estimation and inference. Biostatistics. 24, 635–652 (2023). https://doi.org/10.1093/biostatistics/kxab039

29. Chung, M., Bruno, V.M., Rasko, D.A., Cuomo, C.A., Muñoz, J.F., Livny, J., Shetty, A.C., Mahurkar, A., Dunning Hotopp, J.C.: Best practices on the differential expression analysis of multi-species RNA-seq, (2021)
30. Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., Peddada, S.D.: Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 26, (2015). https://doi.org/10.3402/mehd.v26.27663
31. Kleine Bardenhorst, S., Berger, T., Klawonn, F., Vital, M., Karch, A., Rübsamen, N.: Data Analysis Strategies for Microbiome Studies in Human Populations—a Systematic Review of Current Practice. mSystems. 6, (2021). https://doi.org/10.1128/msystems.01154-20
32. Wallen, Z.D.: Comparison study of differential abundance testing methods using two large Parkinson disease gut microbiome datasets derived from 16S amplicon sequencing. BMC Bioinformatics. 22, (2021). https://doi.org/10.1186/s12859-021-04193-6
33. Yang, L., Chen, J.: A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. Microbiome. 10, (2022). https://doi.org/10.1186/s40168-022-01320-0
34. Hawinkel, S., Mattiello, F., Bijnens, L., Thas, O.: A broken promise: Microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform. 20, 210–221 (2019). https://doi.org/10.1093/bib/bbx104
35. Sequeira, J.C., Rocha, M., Madalena Alves, M., Salvador, A.F.: MOSCA: An automated pipeline for integrated metagenomics and metatranscriptomics data analysis. In: Advances in Intelligent Systems and Computing. pp. 183–191. Springer Verlag (2019)
36. Pereira, M.B., Wallroth, M., Jonsson, V., Kristiansson, E.: Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC Genomics. 19, (2018). https://doi.org/10.1186/s12864-018-4637-6
37. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T.G., Hall, A.B., Lake, K., Landers, C.J., Mallick, H., Plichta, D.R., Prasad, M., Rahnavard, G., Sauk, J., Shungin, D., Vázquez-Baeza, Y., White, R.A., Bishai, J., Bullock, K., Deik, A., Dennis, C., Kaplan, J.L., Khalili, H., McIver, L.J., Moran, C.J., Nguyen, L., Pierce, K.A., Schwager, R., Sirota-Madi, A., Stevens, B.W., Tan, W., ten Hoeve, J.J., Weingart, G., Wilson, R.G., Yajnik, V., Braun, J., Denson, L.A., Jansson, J.K., Knight, R., Kugathasan, S., McGovern, D.P.B., Petrosino, J.F., Stappenbeck, T.S., Winter, H.S., Clish, C.B., Franzosa, E.A., Vlamakis, H., Xavier, R.J., Huttenhower, C.: Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 569, 655–662 (2019). https://doi.org/10.1038/s41586-019-1237-9
38. Peng, X., Li, G., Liu, Z.: Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data. Journal of Computational Biology. 23, 102–110 (2016). https://doi.org/10.1089/cmb.2015.0157

39. Stasinopoulos, D.M., Rigby, R.A.: Journal of Statistical Software Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. (2007)
40. Zupancic, M.L., Cantarel, B.L., Liu, Z., Drabek, E.F., Ryan, K.A., Cirimotich, S., Jones, C., Knight, R., Walters, W.A., Knights, D., Mongodin, E.F., Horenstein, R.B., Mitchell, B.D., Steinle, N., Snitker, S., Shuldiner, A.R., Fraser, C.M.: Analysis of the gut microbiota in the old order amish and its relation to the metabolic syndrome. PLoS One. 7, (2012). https://doi.org/10.1371/journal.pone.0043052
41. Divaris, K., Shungin, D., Rodríguez-Cortés, A., Basta, P. V., Roach, J., Cho, H., Wu, D., Ferreira Zandoná, A.G., Ginnis, J., Ramamoorthy, S., Kinchen, J.M., Kwintkiewicz, J., Butz, N., Ribeiro, A.A., Azcarate-Peril, M.A.: The supragingival biofilm in early childhood caries: Clinical and laboratory protocols and bioinformatics pipelines supporting metagenomics, metatranscriptomics, and metabolomics studies of the oral microbiome. In: Methods in Molecular Biology. pp. 525–548. Humana Press Inc. (2019)
42. Divaris, K., Slade, G.D., Ferreira Zandona, A.G., Preisser, J.S., Ginnis, J., Simancas-Pallares, M.A., Agler, C.S., Shrestha, P., Karhade, D.S., Ribeiro, A. de A., Cho, H., Gu, Y., Meyer, B.D., Joshi, A.R., Azcarate-Peril, M.A., Basta, P. V., Wu, D., North, K.E.: Cohort profile: Zoe 2.0—a community-based genetic epidemiologic study of early childhood oral health. Int J Environ Res Public Health. 17, 1–16 (2020). https://doi.org/10.3390/ijerph17218056
43. McIver, L.J., Abu-Ali, G., Franzosa, E.A., Schwager, R., Morgan, X.C., Waldron, L., Segata, N., Huttenhower, C.: BioBakery: A meta'omic analysis environment. Bioinformatics. 34, 1235–1237 (2018). https://doi.org/10.1093/bioinformatics/btx754
44. Monger, C., Kelly, P.S., Gallagher, C., Clynes, M., Barron, N., Clarke, C.: Towards next generation CHO cell biology: Bioinformatics methods for RNA-Seq-based expression profiling, (2015)
45. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C.: Metagenomic biomarker discovery and explanation. Genome Biol. 12, (2011). https://doi.org/10.1186/gb-2011-12-6-r60
46. Fang, Z., Martin, J., Wang, Z.: Statistical methods for identifying differentially expressed genes in RNA-Seq experiments, (2012)
47. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biol. 11, (2010). https://doi.org/10.1186/gb-2010-11-10-r106
48. Lutz, K.C., Jiang, S., Neugent, M.L., De Nisco, N.J., Zhan, X., Li, Q.: A Survey of Statistical Methods for Microbiome Data Analysis, (2022)
49. Lee, C., Lee, S., Park T: A comparison study of statistical methods for the analysis metagenome data. (2017)