Differential gene expression in metatranscriptomics

Mariana Oliveira, João Carlos Sequeira and Andreia Ferreira Salvador

Centre of Biological Engineering, University of Minho, Campus de Gualtar, Braga, Portugal

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 has 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 relatively limited progress in the development of specialized statistical analysis techniques tailored to this field [3]. Most of the methods 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 or, as it's commonly known, differential gene expression (DGE) 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 Methods

The methods initially employed to identify 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 include 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 – such as edgeR, DESeq2 and voom(+limma). Non-parametric ones include 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 methods will be provided. This review will outline the advantages and limitations of each method. For further statistical information and detailed comparison of these methods, 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].

Commonly used methods that employ the Poisson distribution are **DEGseq** and 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 methods: **baySeq**, **Cuffdiff2**, **DESeq**, **DESeq2**, **EBSeq**, **edgeR**, **NBPseq** and **vst(+limma)**. **BaySeq**, **DESeq** 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 manifested a high sensivity and **bayseq** 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 method 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 Methods for Microbiome Differential Gene Expression Analysis

Parametric Models. Some parametric methods commonly used for DGE that were specifically developed for microbiome data analysis 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 methodology 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]. **ZIBSeq** demonstrates high sensitivity under most scenarios. However, caution is advised in its use, especially when dealing with a high proportion of zeros in small sample sizes, as it may lead to an inflated type I error. Nonetheless, ZIBSeq exhibits favorable properties such as efficient computation time and the ability to handle excess zeros, making it a valuable tool for DE analysis in MT data [3].

Non-Parametric Models. Apart from these parametric approaches, 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, 28]. 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 [29, 30].

Methods Performance. Comprehensive analysis and evaluations of the tools employed in MT DGE have been conducted, yet there has been scant review regarding the methods' performance at the gene or gene-family level [3]. Instead, they commonly evaluate differential abundance or differential expression methods at the species or at the taxon-level [27, 31, 32].

Cho et al. recently conducted a comparative evaluation of statistical analysis methods used in microbial analyses in MT. They assessed recently developed models designed to address zero-inflated over-dispersed counts or compositional data, such as Model-based Analysis of Single-cell Transcriptomics (MAST) and ZIBSeq. The evaluation focused on method 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 Logistic-Beta test (LB), formerly known as 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 and small sample sizes, as mentioned before. 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 tools.

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 recent method was created for an automated and integrated analysis of metagenomics (MG) and MT data performing preprocessing of raw files, assembly, annotation, DGE and comparison of multiple samples [33]. It performs DGE with Bowtie2 (Langmead & Salzberg, 2012), aligning the reads to the scaffolds obtained after the assembly. For MG reads, normalization involves dividing read counts by scaffold length, then further normalizing using EdgeR with the choice of TMM or Relative Log Expression methods (Pereira et al., 2018). MT reads are divided by gene length and normalized similarly. Differential expression analysis utilizes the DESeq2 R package with MT read counts as input. DESeq2 provides *p*-values indicating whether gene expression differs significantly between samples.

Currently, MOSCA uses DESeq2(version 1.81.1) for DGE analysis, generating graphical representations that represent the expression values, such as heatmaps [33]. Meanwhile, it's limited by using only one fold-change value and reporting fold-change for two conditions. Using multiple fold-change values and implementing one-by-one sample comparisons would provide valuable insights into DGE.

4 Objectives

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 achieve this, various methodologies will be compared regarding their statistical foundations and the most suitable one will be selected, thereby enhancing the MOSCA framework.

5 Methodology

To achieve the objectives, a comprehensive review of various approaches used in DGE analysis was conducted (section 2), focusing on systematically assessing the statistical aspects, the strengths and limitations of each methodology within the field of meta-omics analysis.

Through this comprehensive review, the most suitable method for DGE analysis was identified and selected. This decision was made considering the findings of the review, which provided valuable insights into the capabilities and performance of various tools available for DGE analysis [34].

Based on the outcomes of this evaluation, it was determined that replacing DESeq2 with the ZIBSeq in the MOSCA framework would be an interesting change to study and evaluate [34]. 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, 34]. 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 method - 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.

The methodology will involve two main steps. Firstly, the replacement of DESeq2 with ZIBSeq in the MOSCA framework, aiming to improve workflow efficiency and address identified considerations from the review process (Fig. 1). Secondly, the comparison of the performance of the ZIBSeq and DESeq2 using both real and simulated MT datasets. Real datasets will cover various microbial communities and environmental conditions to ensure comprehensive evaluation. Simulated datasets will allow controlled comparisons under different noise levels, fold changes, and sample sizes.

Performance of both tools will be assessed using key metrics such as FDR, sensitivity, variability, specificity a computational efficiency taking into consideration their performance at the gene or gene-family level. Statistical tests will be applied to compare the methods across these metrics.

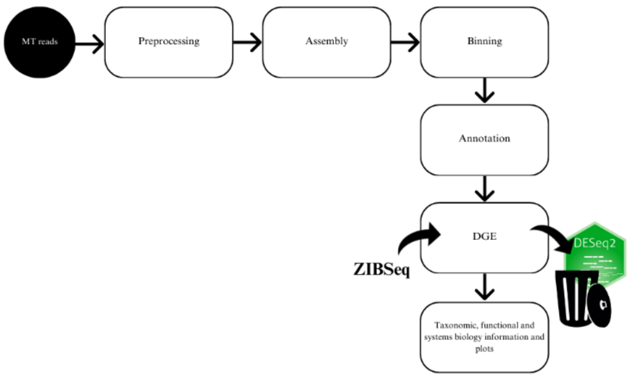

Figure 1. Illustration of MOSCA scheme, with modifications indicated by large arrows.

6 Appendices

Table 1. Differential gene expression analysis methods.

Tool	Normalization	Distribution	Hypothesis test	Dispersion estimation	Ref.
ALDEx2	CLR	Dirichlet-multi-nomial	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 likelihood	[10, 11, 15]

Tool	Normalization	Distribution	Hypothesis test	Dispersion estimation	Ref.
Cuffdiff2	Geometric/FPKM	Beta NB	*t*-test analogical method	Maximum likehood	[13, 20, 35]
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]
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, 29, 36]
MaAsLin	TSS	Gaussian	Wald	Not specified	[4]
metagenomeSeq	CSS	Zero-inflated Gaussian	Moderated *t*	Mixture model with ZIG distribution	[4, 27]

Tool	Normalization	Distribution	Hypothesis test	Dispersion estimation	Ref.
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/Upperquartile	Non-parametric	Ratio of fold change and absolute expression differences	Not specified	[9, 11, 20]
NOIseqBIO	RPKM/TMM/Upperquartile	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]
TSPM	TMM/Not specified	Two-stage Poisson	Likehood ratio/Quasi-likehood approach	Approximated Chi-squared distribution or F-distribution	[10, 17, 37]
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, 38]
ZIBseq	TSS	Zero-inflated beta	Wald/likehood	Maximum likehood	[39, 40]

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. 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

29. 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
30. 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
31. 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
32. 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
33. 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)
34. 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
35. 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)
36. 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
37. Fang, Z., Martin, J., Wang, Z.: Statistical methods for identifying differentially expressed genes in RNA-Seq experiments, (2012)
38. 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
39. 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)
40. Lee, C., Lee, S., Park T: A comparison study of statistical methods for the analysis metagenome data. (2017)
41. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat Methods. 9, 357–359 (2012). https://doi.org/10.1038/NMETH.1923
42. 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)