OMICS A Journal of Integrative Biology
Volume 27, Number 12, 2023

DOI: 10.1089/omi.2023.0232

Commentary

Open camera or QR reader and scan code to access this article and other resources online.

Bridging the Gaps in Meta-Omic Analysis: Workflows and Reproducibility

João Vitor Ferreira Cavalcante,[1] Iara Dantas de Souza,[1] Diego Arthur de Azevedo Morais,[1] and Rodrigo Juliani Siqueira Dalmolin[1,2]

Abstract

The past few years have seen significant advances in the study of complex microbial communities associated with the evolution of sequencing technologies and increasing adoption of whole genome shotgun sequencing methods over the once more traditional Amplicon-based methods. Although these advances have broadened the horizon of meta-omic analyses in planetary health, human health, and ecology from simple sample composition studies to comprehensive taxonomic and metabolic profiles, there are still significant challenges in processing these data. First, there is a widespread lack of standardization in data processing, including software choices and the ease of installing and running attendant software. This can lead to several inconsistencies, making comparing results across studies and reproducing original results difficult. We argue that these drawbacks are especially evident in metatranscriptomic analysis, with most analyses relying on *ad hoc* scripts instead of pipelines implemented in workflow managers. Additional challenges rely on integrating meta-omic data, since methods have to consider the biases in the library preparation and sequencing methods and the technical noise that can arise from it. Here, we critically discuss the current limitations in metagenomics and metatranscriptomics methods with a view to catalyze future innovations in the field of Planetary Health, ecology, and allied fields of life sciences. We highlight possible solutions for these constraints to bring about more standardization, with ease of installation, high performance, and reproducibility as guiding principles.

Keywords: metagenomics, metatranscriptomics, pipelines, reproducibility, sustainable software, data integration

Perspective

Significant advances have been made in microbiology and the study of microbial communities over the past decade. One notable breakthrough is amplicon sequencing, which allows scientists to study the taxonomic composition of an environmental sample. The nascent area of metagenomics was then significantly broadened by the development of whole-metagenome shotgun (WMS) sequencing, which enables the investigation of the full genetic content of samples. This allowed the analysis of functional pathways (Franzosa et al., 2018) as well as the discovery of new microorganisms through metagenome-assembled genomes (Breitwieser et al., 2019). Metagenomics has also contributed to the rise of new fields such as Planetary Health and One Health that view human and ecosystem health intertwined and interdependent, proving to be an impactful approach for integrative areas of study. On the other hand, although WMS data processing software is now widely adopted in the scientific community, there are still multiple challenges in analyzing these data.

Installability, ease of use, and portability among different systems are central challenges in metagenomics and metatranscriptomics software. These difficulties can be mitigated through the use of different methods to improve software, such as code packaging and container technology, as well as toolset curation and workflow management software.

The intrinsic nature of computational science favors reproducibility in the execution of an analysis, as each step can be programmed or automated. However, this usually does

[1]Bioinformatics Multidisciplinary Environment—IMD, Federal University of Rio Grande do Norte, Natal, Brazil.
[2]Department of Biochemistry—CB, Federal University of Rio Grande do Norte, Natal, Brazil.

not happen in practice. This is due to difficulties in software installation, execution, and documentation (Piccolo and Frampton, 2016). Bioinformatics software can be packaged in a multitude of ways, and even though there is no standard on how to best package and share your software, there are some principles that can be easily followed.

For example, prioritizing using a single software version throughout the analysis can boost reproducibility (Nüst et al., 2020). This can be enforced through the use of containerization software, like Docker or Singularity, and there already are plenty of initiatives seeking to standardize bioinformatics software packaging, like BioConda (https://bioconda.github.io/) and BioContainers (https://biocontainers.pro/). These technologies may significantly improve the installability and archival stability of software, leading to an increase in citation rates (Mangul et al., 2019). There is an enhancement in the reproducibility of an analysis and a potential increase in the research impact of it by packaging software and distributing their specific versions with technologies such as Conda, Docker, and Singularity.

In our opinion, no evaluation of metagenomics software has directly ascertained the adoption of software packaging and container technology, although there have been ease-of-use evaluations for metagenomics software (Lindgreen et al., 2016). This could point to a low adoption of these practices in the metagenomics community.

Lindgreen et al. (2016) have also pointed out that toolset choice has a significant impact in metagenomic analysis, not only in terms of computational performance but also in terms of accuracy. Therefore, toolset curation has become common among bioinformaticians, spawning many different scientific workflows for metagenomics data analysis, such as nf-core/mag (Krakau et al., 2022) and MEDUSA (Morais et al., 2022). Moreover, many workflow management software have also been created and adopted by the metagenomics community, primarily Snakemake, Nextflow, and Galaxy.

These tools can greatly enhance reproducibility and ease of use by mitigating platform-specific inaccuracies through their support for isolated task execution. Furthermore, they facilitate the integration of disjointed pieces of code, offering a comprehensive view of the entire analysis pipeline (Wratten et al., 2021). They also provide modularity, allowing users to choose different steps for their analysis. Workflow management software provides significant advantages for metagenomics data processing. However, there is limited adoption of these technologies among bioinformaticians, even among those already involved in toolset curation.

While the limitations discussed here are common to both metagenomics and metatranscriptomics, or, rather, to bioinformatics as a whole, they become more evident with metatranscriptomic methods. Metatranscriptomics helps to provide a clearer understanding of the functional environment in a sample, in a way that is not easily done with metagenomics. For example, metatranscriptomics can show active microbes in a community and which metabolic pathways are most prevalent (Bashiardes et al., 2016), as well as elucidate pathogen–host interactions (Moniruzzaman et al., 2017). Despite its biological potential, there are few performance and accuracy benchmarks for metatranscriptomics pipelines (Shakya et al., 2019).

Additionally, we suggest that many metatranscriptomics pipelines [see Shakya et al. (2019) for an overview] do not adhere to the sustainable software principles mentioned in our present analysis (Table 1). We verified the source code repository and documentation for these pipelines and argue that the application of software packaging, container technology, and workflow management software, practices that enhance the reproducibility and long-term support of these pipelines, are still limited among these tools. We also acknowledge that there are other relevant criteria for their current implementation, such as a focus on web-based analysis, that was not within the scope of this comparison. Furthermore, improvements are evident in recent pipelines, which better tackle the reproducibility issue by implementing container technology such as Docker and Singularity (Taj et al., 2023). Nonetheless, the low adherence to software sustainability practices still shows the necessity for specific and sustainable methods for these types of data.

Methodologies for integrating multiomic data are still few and far between, and sometimes still rely on disconnected scripts without workflow management or containerization. Moreover, benchmarks of omic integration tools are usually restricted to a specific biological question of interest, and, therefore, not generalizable (Subramanian et al., 2020).

Still, there have been notable advances in multiomic data integration, particularly in the Galaxy community, with the development of three integrative meta-omic pipelines that, coupled with a web application, provide an end-to-end

Table 1. Selected Metatranscriptomics Pipelines and Their Comparison in Relation to the Sustainable Software Principles Mentioned in This Paper

Pipeline	*Software packaging or container technology*	*Workflow management*	*Notes*
MetaTrans	N/A	N/A	Website could not be accessed. http://www.metatrans.org
COMAN	N/A	N/A	Web server based.
FMAP	No	No	https://github.com/jiwoongbio/FMAP
SAMSA2	No	No	https://github.com/transcript/samsa2
HUMANn2	Yes (Conda)	No	https://github.com/biobakery/humann
SqueezeMeta	Yes (Conda)	No	https://github.com/jtamames/SqueezeMeta
IMP	Yes (Docker)	Yes (Snakemake)	https://git-r3lab.uni.lu/IMP/IMP
MOSCA	Yes (Conda)	Yes (Snakemake)	https://github.com/iquasere/MOSCA

Each of these pipelines was checked for applications of software packaging, container technology, and use of workflow management software. Access date: October 30, 2023.

N/A, not applicable.

analysis of meta-omic data (Schiml et al., 2023). This could motivate similar developments in other workflow orchestration software communities.

As with most software for bioinformatics, software for analyzing metagenomics and metatranscriptomics data need to be developed with clarity, reproducibility, and reusability as guiding principles. This is further supported when we look at the current metatranscriptomics and data integration ecosystems, these that are still nascent approaches to microbiome studies and therefore still show low adherence to these practices. A wider adoption of software sustainability guidelines in metagenomics and metatranscriptomics methods ensures future research and scientists of the high quality of these methods and their ability to stand the test of time, and it also paves the way for meta-omics to be an indispensable approach to various areas of study, such as ecology and Planetary Health.

Authors' Contributions

J.V.F.C. wrote the original draft and edited it. R.J.S.D. and I.D.d.S. contributed to the discussion of the topic. R.J.S.D., J.V.F.C., and D.A.d.A.M. conceived the original idea. All authors contributed to the drafts and approved the final manuscript.

Author Disclosure Statement

The authors declare there are no conflicting financial interests.

Funding Information

This work was supported by the governmental Brazilian agencies CAPES, grant 88887.834652/2023-00, and CNPq, grant 312305/2021-4.

References

Bashiardes S, Zilberman-Schapira G, Elinav E. Use of metatranscriptomics in microbiome research. Bioinform Biol Insights 2016;10:BBI.S34610; doi: 10.4137/BBI.S34610.

Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform 2019;20(4):1125–1136; doi: 10.1093/bib/bbx120

Franzosa EA, McIver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 2018;15(11):962–968; doi: 10.1038/s41592-018-0176-y

Krakau S, Straub D, Gourlé H, et al. Nf-Core/Mag: A best-practice pipeline for metagenome hybrid assembly and binning. NAR Genom Bioinform 2022;4(1):lqac007; doi: 10.1093/nargab/lqac007

Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep 2016;6(1):19233; doi: 10.1038/srep19233

Mangul S, Mosqueiro T, Abdill RJ, et al. Challenges and recommendations to improve the installability and archival stability of omics computational tools. PLoS Biol 2019;17(6):e3000333; doi: 10.1371/journal.pbio.3000333

Moniruzzaman M, Wurch LL, Alexander H, et al. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nat Commun 2017;8:16054; doi: 10.1038/ncomms16054

Morais DAA, Cavalcante JVF, Monteiro SS, et al. MEDUSA: A pipeline for sensitive taxonomic classification and flexible functional annotation of metagenomic shotgun sequences. Front Genet 2022;13.

Nüst D, Sochat V, Marwick B, et al. Ten simple rules for writing dockerfiles for reproducible data science. PLoS Comput Biol 2020;16(11):e1008316; doi: 10.1371/journal.pcbi.1008316

Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. GigaScience 2016;5(1):30; doi: 10.1186/s13742-016-0135-4

Schiml VC, Delogu F, Kumar P, et al. Integrative meta-omics in galaxy and beyond. Environ Microbiome 2023;18(1):56; doi: 10.1186/s40793-023-00514-9.

Shakya M, Lo C-C, Chain PSG. Advances and challenges in metatranscriptomic analysis. Front Genet 2019;10:904.

Subramanian I, Verma S, Kumar S, et al. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 2020;14:1177932219899051; doi: 10.1177/1177932219899051

Taj B, Adeolu M, Xiong X, et al. MetaPro: A scalable and reproducible data processing and analysis pipeline for metatranscriptomic investigation of microbial communities. Microbiome 2023;11(1):143; doi: 10.1186/s40168-023-01562-6

Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods 2021;18(10):1161–1168; doi: 10.1038/s41592-021-01254-9

Address correspondence to:
Rodrigo Juliani Siqueira Dalmolin, PhD
Bioinformatics Multidisciplinary Environment
Rua do Horto
Lagoa Nova
Natal 59076-550
Brazil

E-mails: rodrigo.dalmolin@imd.ufrn.br;
dalmolin_r@yahoo.com.br

Abbreviation Used

WMS = whole-metagenome shotgun