UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METRÓPOLE DIGITAL
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

JOÃO VITOR FERREIRA CAVALCANTE

O Ecossistema Computacional da Metagenômica: Fluxos de Trabalho e Reprodutibilidade

NATAL - RN
2024

JOÃO VITOR FERREIRA CAVALCANTE

O Ecossistema Computacional da Metagenômica: Fluxos de Trabalho e Reprodutibilidade

Defesa de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal do Rio Grande do Norte.

Área de Concentração: Bioinformática
Linha de Pesquisa: Biologia de Sistemas
Orientador: Rodrigo Juliani Siqueira Dalmolin

NATAL - RN
2024

JOÃO VITOR FERREIRA CAVALCANTE

O Ecossistema Computacional da Metagenômica: Fluxos de Trabalho e Reprodutibilidade

Defesa de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal do Rio Grande do Norte

Área de Concentração: Bioinformática
Linha de Pesquisa: Biologia de Sistemas
Orientador: Rodrigo Juliani Siqueira Dalmolin

Natal, 06 de Dezembro de 2024.

BANCA EXAMINADORA

Prof. Dr. Rodrigo Juliani Siqueira Dalmolin
Universidade Federal do Rio Grande do Norte
(Presidente)

Prof. Dr. Renan Cipriano Moioli
Universidade Federal do Rio Grande do Norte
(Examinador Interno do Programa)

Prof. Dr. Daniel Carlos Ferreira Lanza
Universidade Federal do Rio Grande do Norte
(Examinador Interno do Programa)

Prof. Dr. Marcel da Câmara Ribeiro-Dantas
Universidade Potiguar
(Examinador Externo à Instituição)

Non ut sit repellendus ea sunt dolor.

AGRADECIMENTOS

Nihil ipsum velit cupiditate. Facilis inventore eveniet illo et.
Fugit unde cumque necessitatibus repudiandae non.
Voluptatem hic a numquam enim aperiam voluptatem.

“Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.”
(Richard P. Feynman)

RESUMO

No resumo são ressaltados o objetivo da pesquisa, o método utilizado, as discussões e os resultados com destaque apenas para os pontos principais. O resumo deve ser significativo, composto de uma sequência de frases concisas, afirmativas, e não de uma enumeração de tópicos. Não deve conter citações. Deve usar o verbo na voz ativa e na terceira pessoa do singular. O texto do resumo deve ser digitado, em um único bloco, sem espaço de parágrafo. O espaçamento entre linhas é simples e o tamanho da fonte é 12. Abaixo do resumo, informar as palavras-chave (palavras ou expressões significativas retiradas do texto) ou, termos retirados de thesaurus da área. Deve conter de 150 a 500 palavras. O resumo é elaborado de acordo com a NBR 6028.

Palavras-chave: Palavra-chave 1. Palavra-chave 2.

ABSTRACT

Resumo traduzido para outros idiomas, neste caso, inglês. Segue o formato do resumo feito na língua vernácula. As palavras-chave traduzidas, versão em língua estrangeira, são colocadas abaixo do texto precedidas pela expressão “Keywords”, separadas por ponto.

Keywords: Keyword 1. Keyword 2.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

16S	sequenciamento da subunidade ribossomal 16S de genomas bacterianos
ASV	variantes de sequência amplicon - do inglês *amplicon sequence variant*
DNA	Ácido Desoxirribonucléico
MS	metagenômica *shotgun*
OTU	unidades taxonômicas operacionais - do inglês *operational taxonomic units*

SUMÁRIO

1 INTRODUÇÃO

1.1 VISÃO GERAL - METAGENÔMICA

A história da vida microscópica, ou microbiana, no planeta Terra supera a história da vida macroscópica por milhares de anos (MAGNABOSCO *et al.*, 2024). A metagenômica surge como uma abordagem que possibilita descobertas acerca da vida microbiana através do sequenciamento genético. Avanços no que viria eventualmente a se tornar a metagenômica surgem ainda nos anos 90, com o primeiro sequenciamento de genoma completo de um organismo de vida livre, a bactéria *Haemophilus influenza* (WOOLEY; GODZIK; FRIEDBERG, 2010). Esse ponto na história científica marca o primeiro uso bem sucedido do que vem a ser chamado de *whole-genome shotgun*, ou sequenciamento de genoma completo, no qual a amostra possui seu conteúdo genético fragmentado em *reads*, ou leituras, que são então sequenciadas. Essa técnica viria a ser refinada e aplicada para amostras ambientais, seja este ambiente uma amostra de solo florestal ou uma biópsia intestinal. Nessa nova técnica se buscou sequenciar o conteúdo genético que compreenda os diferentes microorganismos presentes em tais amostras, originando assim o que será aqui descrito como metagenômica *shotgun* (MS).

No entanto, o estudo de comunidades microbianas se populariza de fato com uma técnica que não busca capturar o conteúdo genético total de uma amostra, mas apenas uma subregião de seu Ácido Desoxirribonucléico (DNA) ribossomal que possua ao mesmo tempo regiões conservadas, capaz de serem passíveis de anelamento por *primers*, e regiões hipervariáveis, capazes de distinguir um microorganismo de outro. Em bactérias o ribotipo selecionado foi o DNA que codifica a subunidade 16S, que é amplificada e então sequenciada. O sequenciamento da subunidade ribossomal 16S de genomas bacterianos (16S), também denominado metataxonômica (MARCHESI; RAVEL, 2015), possibilitou uma maneira simples, e pouco computacionalmente intensiva quando comparada à MS (TREMBLAY; SCHREIBER; GREER, 2022), para realizar identificação de táxons bacterianos em uma amostra ambiental. Ademais, técnicas computacionais posteriores ultimamente facilitariam a conexão de informação funcional às abundâncias taxonômicas obtidas através desta técnica.

Dessa maneira, possuímos atualmente, duas possíveis abordagens para se estudar comunidades microbianas, a MS e o 16S. Essas abordagens, sobretudo a MS, geram um alto volume de dados, e portanto, há a necessidade do desenvolvimento de ferramentas computacionais que processem esses dados e gerem informação científica que descreva de forma acurada e reproduzível achados acerca do microambiente do estudo (COMIN *et al.*, 2021).

1.2 DESENVOLVIMENTO DE SOFTWARE CIENTÍFICO

Metodologias computacionais constituem parte indissociável da biologia molecular moderna, com novas ferramentas, abordagens e fluxos de trabalho, isto é, metodologias que agregam diversas ferramentas em sequência, surgindo a cada momento. Nesse contexto, muito tem se discutido acerca da qualidade do software desenvolvido, não apenas do ponto de vista de qualidade científica ou estatística da análise a ser realizada, mas também qualidade a partir de um ponto de vista técnico.

Tipicamente software é tratado como um ponto adjacente à uma análise científica, e não seu principal produto. Parcialmente isto se dá devido ao modelo de desenvolvimento de software científico atualmente vigente, que se dá por meio de projetos de doutorado e mestrado, dificultando, dessa maneira, manutenção a longo prazo (ALTSCHUL *et al.*, 2013) (MANGUL; MARTIN *et al.*, 2019). Portanto, produzir metodologias que sejam usáveis a longo prazo sem necessitarem de alta manutenção é essencial. Nesse sentido, a utilização de tecnologias que facilitem a instalação de software, como gerenciadores de ambiente, ou tecnologias capazes de encapsular um ambiente computacional, como contêineres, são indispensáveis para a garantia de reprodutibilidade de qualquer método computacional, por garantirem o isolamento das dependências de um *software* indefinidamente (KADRI *et al.*, 2022).

No entanto, garantir a capacidade de instalação do software é um ponto basal para determinar a qualidade de código científico. Outros princípios, aqui denominados como boas práticas de desenvolvimento de software científicos, são igualmente indispensáveis para que uma metodologia seja utilizável a longo prazo, além de capaz de gerar resultados científicos interpretáveis e consistentes. Além da instalabilidade, alguns outros princípios que garantem a qualidade de um software a longo prazo são uma documentação descritiva, com exemplos práticos de utilização, suporte multiplataforma (MANGUL; MOSQUEIRO *et al.*, 2019) e, para softwares de processamento de dados, relatórios interativos ou logs acessíveis (PERKEL, 2018). A implementação de tais princípios em um *software* científico não apenas aumenta sua usabilidade, mas também aumenta as citações de seus artigos, aumentando, consequentemente, o alcance dessas metodologias (MANGUL; MARTIN *et al.*, 2019).

1.3 O ECOSSISTEMA COMPUTACIONAL EM METAGENÔMICA

As metodologias de processamento de dados 16S em grande parte já estão estabelecidas, dada a idade mais avançada da abordagem. Nesse sentido, vemos que o cerne das abordagens trata de atribuir identificadores únicos às sequências 16S obtidas, caracterizando assim táxons distintos. Esses identificadores podem ser atribuídos através de métodos de agrupamento, caracterizando as abordagens baseadas em unidades taxonômicas operacionais - do inglês *operational taxonomic units* (OTU),

que agrupam sequências com pelo menos 97% de identidade em grupos biológicos distintos, ou abordagens baseadas em variantes de sequência amplicon - do inglês *amplicon sequence variant* (ASV), que, através de modelos estatísticos, tentam definir variações biológicas reais na sequência - contrastadas com variações devido a erros de sequenciamento, e dessa maneira obter uma resolução taxonômica maior comparada às baseadas em OTU (CHIARELLO *et al.*, 2022). Nesse contexto, observamos metodologias que buscam, a partir dos OTU ou ASV identificados, inferir abundâncias de vias metabólicas específicas, tirando proveito de bancos de dados de informação curada a respeito desses táxons que possua mapeamento das famílias gênicas de seu genoma a funções biológicas (DOUGLAS *et al.*, 2020). Quanto a fluxos de trabalho para processamento de dados 16S, destaca-se o *nf-core/ampliseq* (STRAUB *et al.*, 2020), que implementa várias das boas práticas de software citadas anteriormente, como suporte multi-plataforma, documentação descritiva e exemplificada e relatórios automáticos interativos.

Por outro lado, o campo do desenvolvimento de metodologias para dados de MS ainda é bastante fértil, com novas técnicas computacionais desenvolvidas rotineiramente (LIU *et al.*, 2021). De forma geral, podemos agrupar as metodologias em duas grandes categorias: Metodologias livres de montagem, isto é, aquelas que se utilizam apenas da informação contidas nas leituras para obter seus resultados, e metodologias baseadas em montagem, que primeiro realizam a montagem de leituras em sequências contíguas (ou *contigs*), que fornecerá então a base para o processamento seguinte (BREITWIESER; LU; SALZBERG, 2019). Apesar da capacidade que métodos baseados em montagem tem de descobrir novos organismos e montar genomas inéditos, métodos livres de montagem apresentam certas vantagens, sobretudo quando consideramos dados com baixa cobertura de sequência, o que pode resultar em montagens pouco precisas (AYLING; CLARK; LEGGETT, 2020).

No que se diz respeito a essas metodologias, vemos que as baseadas em montagem são amplas e cobrem os mais diversos aspectos do processamento de dados de MS, com exemplos como nf-core/mag (KRAKAU *et al.*, 2022) e metaphor (SALAZAR *et al.*, 2023). No entanto, quando observamos métodos livres de montagem, vemos um cenário mais escasso, sobretudo quando consideramos apenas fluxos de trabalho, ou *pipelines*, orquestrados por linguagems de gerenciamento de metodologias científicas, como Nextflow (DI TOMMASO *et al.*, 2017) ou Snakemake (MÖLDER *et al.*, 2021). No contexto de métodos livres de montagem para dados MS, vale ressaltar o fluxo de trabalho MEDUSA (MORAIS *et al.*, 2022), que apresentou boa sensitividade e flexibilidade para análises de classificação taxonômica e anotação funcional.

Nesse sentido, há a necessidade de desenvolver uma metodologia para dados de MS que siga boas práticas de desenvolvimento de software científico e que tenha como princípios norteadores a reprodutibilidade, documentação descritiva e prática e

interpretabilidade. Este último ponto que se torna especialmente relevante ao considerarmos a complexidade e a alta dimensionalidade desses dados.

2 OBJETIVOS

2.1 GERAL

Obter uma visão geral do ecossistema computacional em metagenômica atual e como ele se associa com princípios de desenvolvimento de software científico, desenvolvendo então uma metodologia para dados de MS que possibilite uma análisa metagenômica compreensiva, de forma reprodutível e flexível.

2.2 ESPECÍFICOS

- Avaliar o atual ferramentário computacional para dados de metagenômica e sua adesão a princípios de desenvolvimento de software sustentável.
- Desenvolver uma metodologia robusta, flexível e acessível para análise de dados de MS.

CAPÍTULO 1

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

Escrito por: João Vitor Ferreira Cavalcante, Iara Dantas de Souza, Diego Arthur de Azevedo Morais e Rodrigo Juliani Siqueira Dalmolin

Artigo publicado no periódico OMICS: A Journal of Integrative Biology

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

CAPÍTULO 2

Artigo: EURYALE: A versatile Nextflow pipeline for taxonomic classification and functional annotation of metagenomics data

Escrito por: João Vitor Ferreira Cavalcante, Iara Dantas de Souza, Diego Arthur de Azevedo Morais e Rodrigo Juliani Siqueira Dalmolin

Artigo submetido ao periódico IEEE Access

EURYALE: A versatile Nextflow pipeline for taxonomic classification and functional annotation of metagenomics data

João Vitor F. Cavalcante
Bioinformatics Multidisciplinary Environment
Federal University of Rio Grande do Norte
Natal, Brazil
0000-0001-7513-7376

Iara Dantas de Souza
Bioinformatics Multidisciplinary Environment
Federal University of Rio Grande do Norte
Natal, Brazil
0000-0002-2550-6150

Diego A. A. Morais
Bioinformatics Multidisciplinary Environment
Federal University of Rio Grande do Norte
Natal, Brazil
0000-0002-7357-3446

Rodrigo J. S. Dalmolin
Department of Biochemistry
Federal University of Rio Grande do Norte
Natal, Brazil
0000-0002-1688-6155

***Abstract*—EURYALE is a Nextflow pipeline designed for the sensitive taxonomic classification and flexible functional annotation of metagenomic shotgun sequences. It provides a comprehensive solution for preprocessing, assembly, alignment, taxonomic classification, and functional annotation of metagenomic data. EURYALE builds upon the Snakemake-based pipeline MEDUSA. EURYALE inherits the tools present in MEDUSA, selected based on rigorous benchmarks for performance, accuracy, and sensitivity. The new pipeline has been developed with the nf-core pipeline template, which focuses on modularity, allowing a high degree of parameterization. EURYALE provides easier resource management, enforcing strict memory and CPU requirements based on its Nextflow configuration. It has also become more versatile, as it can be executed using Docker and Singularity, which further extends its usability across various platforms. It can also natively take advantage of computational infrastructures such as SLURM and Amazon Web Services. EURYALE inherits the sensitivity in taxonomic classification and flexibility in functional annotation of its predecessor, combined with improved versatility.**

***Index Terms*—Metagenomic Analysis; Nextflow; Taxonomic Classification; Functional Annotation; Pipeline**

I. Introduction

The growth of metagenomics in the last few years as the main approach to study complex microbial communities is apparent, with many different methodologies arising to process whole-metagenome shotgun (WMS) data. Metagenomics consists of an approach to sequence the genetic content of an environmental sample, be that environment a host-associated tissue, like the human gut, be it something like a soil sample.

Metagenomics has typically been restricted, in the past, to amplicon sequencing, an approach nowadays referred to as 'Metataxonomics' [1], which allows researchers to investigate the taxonomic composition of a sample by sequencing specific genomic regions, such as the 16S region in bacteria. WMS, on the other hand, provides not only these specific regions, but a more representative genetic content of a sample, allowing for bacterial assembly, as well as investigations of functional information.

WMS data, due to its size and complexity, presents a series of challenges in regards to its processing. The main challenges that methodologies aim to tackle are: Performance issues and data storage, as metagenomic datasets can be large and unwieldy [2], which can be improved by making metagenomic pipelines usable in cloud environments [3]; Sequence contamination from other sources [4]; And useful and easy-to-parse results, given the inherent complexity of these data [5].

The methods to process WMS data can be broadly divided into two: Assembly-free methods, i.e., those that rely on direct read classification; and assembly-based methods, i.e., methods that first perform read assembly into contigs prior to classification and annotation [6].

Assembly-free methods provide an advantage to assembly-based methods particularly regarding performance, but advantages have also been noted in using these methods with low coverage data, where assembly-based ones can often lead to inaccurate results [7]. Additionally, assembly-free methods have been used to study the biodiversity in marine environments [8] as well as compositional differences in major depressive disorder [9], showing its potential to empower scientific discovery.

Among metagenomics methods, most have organized themselves into pipelines orchestrated by workflow managers, such as Nextflow [10] and Snakemake [11], which increase modularization, parameterization, portability and ease of installation for these softwares [12][13]. Although pipelines covering most of the metagenomic analysis process through assembly-

979-8-3503-5663-2/24/$31.00

based strategies, such as nf-core/mag [14], Metaphor [15] and MGnify [16], are widespread, assembly-free methods are under explored and pipelines are few and far between.

One highlight among assembly-free methods is the MEDUSA[1] pipeline [18], which was built based on careful benchmarking of multiple WMS data analysis tools, which were finally brought together in a pipeline orchestrated by Snakemake. Although MEDUSA proved to be a great advance in this field, there are some implementation details that could be improved. First, portability was limited: MEDUSA could only run through a BioConda-based environment [19], which is error-prone and less stable in the long term than a Docker or Singularity image [20]. Secondly, MEDUSA's DSL of choice, Snakemake, has proven hard to work with and maintain, which is something that others in the field of workflow development have observed [21][22][20], with most preferring Nextflow. Lastly, MEDUSA implemented hard-coded references in its code which were bothersome to change, requiring the user to directly alter the pipeline's source code.

In this context, we re-implemented MEDUSA's software in a new, Nextflow-orchestrated pipeline, called EURYALE, which is more portable, providing dedicated Docker and Singularity images, as well as Nextflow's native support of different HPC schedulers. It is also more parameterizable, removing the need for direct source code modifications. Euryale, as in Greek mythology, is the elder, immortal sister of Medusa, and this name represents the long-term support and stability we aim to provide with this new pipeline. Here we present the decisions we took in developing EURYALE and the results that can be acquired from it.

II. Methods

EURYALE was implemented using the nf-core [23] pipeline template, with each software included as part of the pipeline made available through BioConda [19] and BioContainers [24], enabling execution through conda, Docker and Singularity. The nf-core pipeline template and NextFlow enable high parameterization and customizable resource allocation, while also integrating well with High Performance Computing (HPC) schedulers, such as SLURM (https://slurm.schedmd.com/), as well as cloud environments, such as Amazon Web Services.

MicroView, included as a reporting tool in EURYALE, is a tool implemented in the Python programming language. It creates static HTML reports containing diversity information for a taxonomic profiling sample resulting from Kraken 2 [25] or Kaiju [26]. In each file, only the taxonomically classified reads (i.e., the reads properly attributed to a known taxon) are considered for diversity calculations. The abundance corresponds to the number of reads attributed to a given taxon. The diversity measures used, i.e. the Shannon index and Bray-Curtis distances, were computed using these read quantifications through the `alpha_diversity` and `beta_diversity` functions provided by SciKit-bio [27].

[1]Distinct and separate from MEDUSA as described by F. H. Karlsson, I. Nookaew, and J. Nielsen [17].

III. Results and discussion

A. Implementation details

MEDUSA's tools were fully reimplemented in EURYALE (Fig. 1), including custom solutions implemented in the original MEDUSA like the annotate package for functional annotation. EURYALE consists of 5 general steps or "subworkflows", the first of which comprises general read preprocessing. This step is then followed by host read removal, in the case of samples coming from hosts with known reference genomes, such as human microbiome data. We can then perform an optional - and disabled by default - assembly step, which comes prior to the taxonomic classification and functional annotation. Every step can be skipped, ensuring high customization to EURYALE's users.

Both MEDUSA and EURYALE have as main inputs the FASTQ files containing reads, as well as an assortment of reference databases, for both the taxonomic classification and the functional annotation. While MEDUSA downloaded the references which were directly defined in its source code, EURYALE relies exclusively on user defined references, allowing for a more fine-grained customization. Additionally, EURYALE provides a separate workflow to download the same set of references as MEDUSA's, in case full compatibility with previous results is required.

Below we provide an example command for executing the 'download' entry for EURYALE, which acquires these references. The different parameters select which references will be downloaded in this execution. Further details are explained in the EURYALE documentation: https://dalmolingroup.github.io/euryale/.

```
nextflow run dalmolingroup/euryale \
  --download_functional \
  --download_kaiju \
  --download_host \
  --outdir <output directory> \
  -entry download \
  -profile docker
```

EURYALE provides a new option for taxonomic classification, Kraken 2 [25], which was previously restricted to Kaiju [26]. This decision was taken in view of the Kraken's team active development and high adoption by the metagenomics community, seen by the recent benchmarks which point to better precision and sensitivity when compared to other softwares[28][29][25].

One departure from MEDUSA was to remove steps that downloaded reference data, since these were prone to failure and required the user to manually alter source code in case of changing the reference databases to be used. Now EURYALE has over 20 parameters that can be added and modified in each execution by simply changing them on the command line, guaranteeing better adaptation to distinct use cases. By taking advantage of the nf-core infrastructure and template for creating pipelines [23], EURYALE also provides a high degree of parameterization for resource allocation, with different

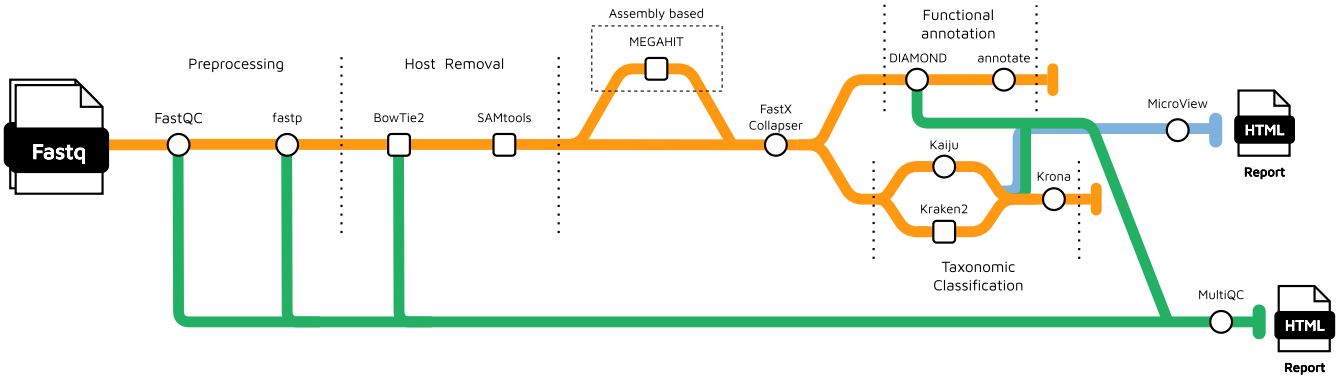


Fig. 1. Diagram showing each step of the EURYALE pipeline, starting with FastQ reads as input. Circles represent steps that are enabled by default, while squares represent steps that are optional or disabled by default.

process labels that specify the number of CPUs and amount of RAM to be used by each process, which are strictly enforced when using cloud environments or HPC schedulers.

Below we provide an example command for executing all of EURYALE's subworkflows in a Docker environment. The input to the pipeline is all samples, one-per-line, in a comma-separated table with sample name being the first column and the other two being paired FastQ files. The pipeline needs, by default, references for the Kaiju database - or Kraken 2 if you chose that taxonomy profiler - the FASTA reference file for DIAMOND [30], a FASTA reference for the host genome, in case the host read removal subworkflow is enabled, and an ID mapping file between the gene IDs in the reference and the desired functional database.

```
nextflow run dalmolingroup/euryale \
  --input samplesheet.csv \
  --outdir <output directory> \
  --kaiju_db <kaiju database> \
  --reference_fasta <reference FASTA> \
  --host_fasta <host reference FASTA> \
  --id_mapping <ID mapping file> \
  -profile docker
```

B. Automated Reporting

One important change from MEDUSA to EURYALE is the enhancements regarding automated report creation and information interchange between the pipeline and its end user. In MEDUSA, we started working towards this with per-sample Krona [31] visualizations that provide a simple overview over the proportion of each species' presence in each sample. Now, in EURYALE, apart from these same reports, we have also implemented two more: MultiQC [32] reports that show general quality control metrics and other measurements throughout various processes and MicroView reports, which have been customly developed for dealing with the taxonomic classification data coming from the results of this pipeline.

MultiQC reports contain a brief overview of nearly every tool included as part of EURYALE, containing visualizations regarding sequence preprocessing (Fig. 2A), alignment and taxonomic classification (Fig. 2B). Visualizations such as these can both serve as quality control measures as well as indicate which taxa are more abundant in your samples, potentially generating insights about microbial biomarkers.

Apart from this, EURYALE also implements a new tool for reporting: MicroView. This new tool is focused towards calculating common diversity metrics based on the read classification data and providing simple, but insightful, visualizations about these metrics. Currently supported in this version of EURYALE are visualizations for Bray-Curtis beta-diversity, which is used to plot a Principal Coordinate Analysis, the Shannon index and Pielou evenness (Fig 3), making exploratory biodiversity analyses easily available, which could point to differences in biodiversity or species' evenness among different groups of samples, or the general distribution of these measures, as the figure illustrates.

We see these automated reports as a strong improvement from the previous version, as they increase the pipeline's usability, giving general directions on how to better explore your data and facilitating the generation of new research questions.

IV. Conclusions

EURYALE, a novel metagenomics pipeline, bases itself upon previous well-curated pipelines like MEDUSA, taking its sensitivity, while improving its versatility. The pipeline adds support for container technology through Docker and Singularity; Native support for different HPC schedules as well as cloud environments; And more parameterization, allowing the user to choose different references and select specific pipeline steps to run. These changes have the potential to solve many issues common to WMS data analysis. Furthermore, EURYALE builds upon MEDUSA by adding new automated reports, providing metrics on data quality control, taxonomic classification and diversity, enhancing the methodology's usability as well as empowering its users to quickly generate new scientific insights. Overall, EURYALE, a novel Nextflow pipeline for WMS data analysis, advances metagenomics data

processing by tackling some of its main issues, particularly regarding computational versatility and interpretability of results.

Code Availability Statement

EURYALE is available through a GitHub repository, which can be found at https://github.com/dalmolingroup/euryale, with documentation available through https://dalmolingroup.github.io/euryale/. MicroView, although executed as part of EURYALE, can be executed stand-alone and has its own source code available in the following repository: https://github.com/dalmolingroup/microview.

Acknowledgments

This work was supported by the governmental Brazilian agencies CAPES, grant 88887.834652/2023-00, and CNPq, grant 312305/2021-4. We would also like to thank NPAD / UFRN for computational resources.

References

[1] J. R. Marchesi and J. Ravel, "The vocabulary of microbiome research: A proposal," *Microbiome*, vol. 3, no. 1, p. 31, Jul. 2015, ISSN: 2049-2618. DOI: 10.1186/s40168-015-0094-5. (visited on 06/19/2023).

[2] S. Hunter, M. Corbett, H. Denise, *et al.*, "EBI metagenomics—a new resource for the analysis and archiving of metagenomic data," *Nucleic Acids Research*, vol. 42, no. D1, pp. D600–D606, Jan. 2014, ISSN: 0305-1048. DOI: 10.1093/nar/gkt961. (visited on 06/01/2024).

[3] D. D'Agostino, L. Morganti, E. Corni, D. Cesini, and I. Merelli, "Combining Edge and Cloud computing for low-power, cost-effective metagenomics analysis," *Future Generation Computer Systems*, vol. 90, pp. 79–85, Jan. 2019, ISSN: 0167-739X. DOI: 10.1016/j.future.2018.07.036. (visited on 06/02/2024).

[4] I. Laudadio, V. Fulci, L. Stronati, and C. Carissimi, "Next-Generation Metagenomics: Methodological Challenges and Opportunities," *OMICS: A Journal of Integrative Biology*, vol. 23, no. 7, pp. 327–333, Jul. 2019. DOI: 10.1089/omi.2019.0073. (visited on 06/01/2024).

[5] P. ten Hoopen, R. D. Finn, L. A. Bongo, *et al.*, "The metagenomic data life-cycle: Standards and best practices," *GigaScience*, vol. 6, no. 8, gix047, Aug. 2017, ISSN: 2047-217X. DOI: 10.1093/gigascience/gix047. (visited on 06/01/2024).

[6] F. P. Breitwieser, J. Lu, and S. L. Salzberg, "A review of methods and databases for metagenomic classification and assembly," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1125–1136, Jul. 2019, ISSN: 1477-4054. DOI: 10.1093/bib/bbx120.

[7] M. Ayling, M. D. Clark, and R. M. Leggett, "New approaches for metagenome assembly with short reads," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 584–594, Mar. 2020, ISSN: 1477-4054. DOI: 10.1093/bib/bbz020. (visited on 06/01/2024).

[8] B. C. F. Santiago, I. D. de Souza, J. V. F. Cavalcante, *et al.*, "Metagenomic Analyses Reveal the Influence of Depth Layers on Marine Biodiversity on Tropical and Subtropical Regions," *Microorganisms*, vol. 11, no. 7, p. 1668, Jul. 2023, ISSN: 2076-2607. DOI: 10.3390/microorganisms11071668. (visited on 06/27/2023).

[9] J. Mayneris-Perxachs, A. Castells-Nobau, M. Arnoriaga-Rodríguez, *et al.*, "Microbiota alterations in proline metabolism impact depression," *Cell Metabolism*, vol. 34, no. 5, 681–701.e10, May 2022, ISSN: 1550-4131. DOI: 10.1016/j.cmet.2022.04.001. (visited on 06/01/2024).

[10] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature Biotechnology*, vol. 35, no. 4, pp. 316–319, Apr. 2017, ISSN: 1546-1696. DOI: 10.1038/nbt.3820. (visited on 06/12/2023).

[11] F. Mölder, K. P. Jablonski, B. Letcher, *et al.*, "Sustainable data analysis with Snakemake," *F1000Research*, vol. 10, p. 33, Apr. 2021, ISSN: 2046-1402. DOI: 10.12688/f1000research.29032.2. (visited on 06/12/2023).

[12] J. V. F. Cavalcante, I. D. de Souza, D. A. d. A. Morais, and R. J. S. Dalmolin, "Bridging the Gaps in Meta-Omic Analysis: Workflows and Reproducibility," *OMICS: A Journal of Integrative Biology*, Nov. 2023. DOI: 10.1089/omi.2023.0232. (visited on 12/02/2023).

[13] L. Wratten, A. Wilm, and J. Göke, "Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers," *Nature Methods*, vol. 18, no. 10, pp. 1161–1168, Oct. 2021, ISSN: 1548-7105. DOI: 10.1038/s41592-021-01254-9. (visited on 06/12/2023).

[14] S. Krakau, D. Straub, H. Gourlé, G. Gabernet, and S. Nahnsen, "Nf-core/mag: A best-practice pipeline for metagenome hybrid assembly and binning," *NAR Genomics and Bioinformatics*, vol. 4, no. 1, lqac007, Mar. 2022, ISSN: 2631-9268. DOI: 10.1093/nargab/lqac007. (visited on 04/04/2023).

[15] V. W. Salazar, B. Shaban, M. d. M. Quiroga, *et al.*, "Metaphor—A workflow for streamlined assembly and binning of metagenomes," *GigaScience*, vol. 12, giad055, Jan. 2023, ISSN: 2047-217X. DOI: 10.1093/gigascience/giad055. (visited on 03/26/2024).

[16] T. A. Gurbich, A. Almeida, M. Beracochea, *et al.*, "MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues," *Journal of Molecular Biology*, Computation Resources for Molecular Biology, vol. 435, no. 14, p. 168016, Jul. 2023, ISSN: 0022-2836. DOI: 10.1016/j.jmb.2023.168016. (visited on 03/23/2024).

[17] F. H. Karlsson, I. Nookaew, and J. Nielsen, "Metagenomic Data Utilization and Analysis (MEDUSA) and Construction of a Global Gut Microbial Gene Catalogue," *PLOS Computational Biology*, vol. 10, no. 7,

e1003706, 10 de jul. de 2014, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003706. (visited on 06/01/2024).

[18] D. A. A. Morais, J. V. F. Cavalcante, S. S. Monteiro, M. A. B. Pasquali, and R. J. S. Dalmolin, "MEDUSA: A Pipeline for Sensitive Taxonomic Classification and Flexible Functional Annotation of Metagenomic Shotgun Sequences," *Frontiers in Genetics*, vol. 13, 2022, ISSN: 1664-8021. (visited on 07/10/2023).

[19] B. Grüning, R. Dale, A. Sjödin, *et al.*, "Bioconda: Sustainable and comprehensive software distribution for the life sciences," *Nature Methods*, vol. 15, no. 7, pp. 475–476, Jul. 2018, ISSN: 1548-7105. DOI: 10.1038/s41592-018-0046-7. (visited on 06/12/2023).

[20] S. Grayson, D. Marinov, D. S. Katz, and R. Milewicz, "Automatic Reproduction of Workflows in the Snakemake Workflow Catalog and nf-core Registries," in *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*, ser. ACM REP '23, New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 74–84, ISBN: 9798400701764. DOI: 10.1145/3589806.3600037. (visited on 07/03/2023).

[21] M. Jackson, K. Kavoussanakis, and E. W. J. Wallace, "Using prototyping to choose a bioinformatics workflow management system," *PLOS Computational Biology*, vol. 17, no. 2, e1008622, 25 de fev. de 2021, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008622. (visited on 03/23/2024).

[22] F. M. Celebi, E. McDaniel, and T. Reiter, "Creating reproducible workflows for complex computational pipelines," *Arcadia Science*, Mar. 2023. DOI: 10.57844/arcadia-cc5j-a519. (visited on 04/04/2023).

[23] P. A. Ewels, A. Peltzer, S. Fillinger, *et al.*, "The nf-core framework for community-curated bioinformatics pipelines," *Nature Biotechnology*, vol. 38, no. 3, pp. 276–278, Mar. 2020, ISSN: 1546-1696. DOI: 10.1038/s41587-020-0439-x. (visited on 06/12/2023).

[24] F. da Veiga Leprevost, B. A. Grüning, S. Alves Aflitos, *et al.*, "BioContainers: An open-source and community-driven framework for software standardization," *Bioinformatics*, vol. 33, no. 16, pp. 2580–2582, Aug. 2017, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx192. (visited on 07/10/2023).

[25] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2," *Genome Biology*, vol. 20, no. 1, p. 257, Nov. 2019, ISSN: 1474-760X. DOI: 10.1186/s13059-019-1891-0. (visited on 03/24/2024).

[26] P. Menzel, K. L. Ng, and A. Krogh, "Fast and sensitive taxonomic classification for metagenomics with Kaiju," *Nature Communications*, vol. 7, no. 1, p. 11 257, Apr. 2016, ISSN: 2041-1723. DOI: 10.1038/ncomms11257. (visited on 03/24/2024).

[27] J. R. Rideout, G. Caporaso, E. Bolyen, *et al.*, *Biocore/scikit-bio: Scikit-bio 0.5.9: Maintenance release*, Zenodo, Aug. 2023. DOI: 10.5281/zenodo.8209901. (visited on 03/26/2024).

[28] A. R. Odom, T. Faits, E. Castro-Nallar, K. A. Crandall, and W. E. Johnson, "Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data," *Scientific Reports*, vol. 13, no. 1, p. 13 957, Aug. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-40799-x. (visited on 03/26/2024).

[29] F. Jurado-Rueda, L. Alonso-Guirado, T. E. Perea-Chamblee, *et al.*, "Benchmarking of microbiome detection tools on RNA-seq synthetic databases according to diverse conditions," *Bioinformatics Advances*, vol. 3, no. 1, vbad014, Jan. 2023, ISSN: 2635-0041. DOI: 10.1093/bioadv/vbad014. (visited on 03/26/2024).

[30] B. Buchfink, K. Reuter, and H.-G. Drost, "Sensitive protein alignments at tree-of-life scale using DIAMOND," *Nature Methods*, vol. 18, no. 4, pp. 366–368, Apr. 2021, ISSN: 1548-7105. DOI: 10.1038/s41592-021-01101-x. (visited on 03/27/2024).

[31] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a Web browser," *BMC Bioinformatics*, vol. 12, no. 1, p. 385, Sep. 2011, ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-385. (visited on 03/25/2024).

[32] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: Summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw354. (visited on 03/12/2024).

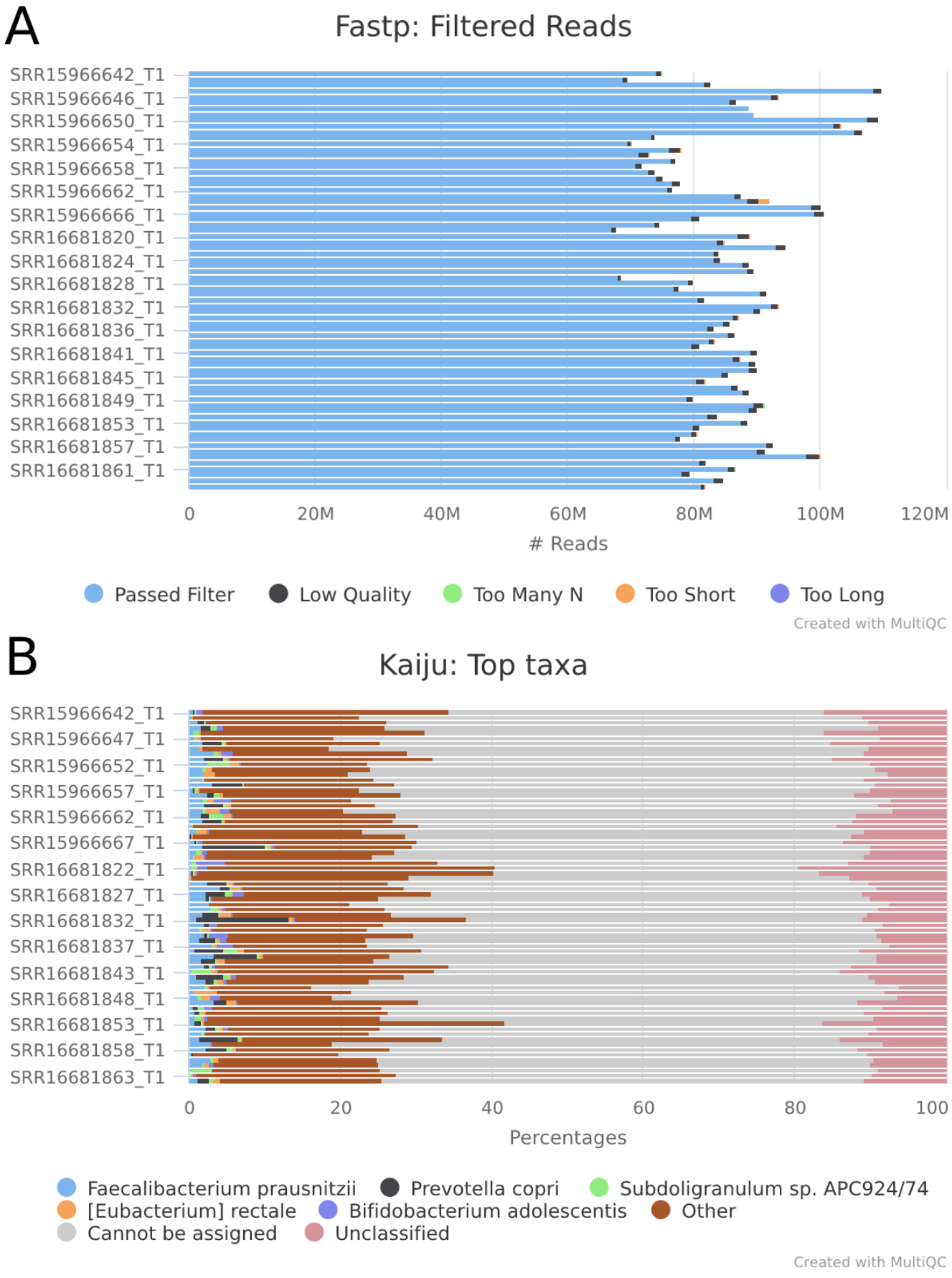


Fig. 2. Example figures included as part of EURYALE's MultiQC report. **A**: Figure for the quality control section, showing the number of reads filtered out in each sample by fastp and their respective thresholds for filtering. **B**: Figure for the taxonomic classification section, showing the percentage of reads assigned to different taxa in each sample by Kaiju. In the HTML report the figures are interactive.

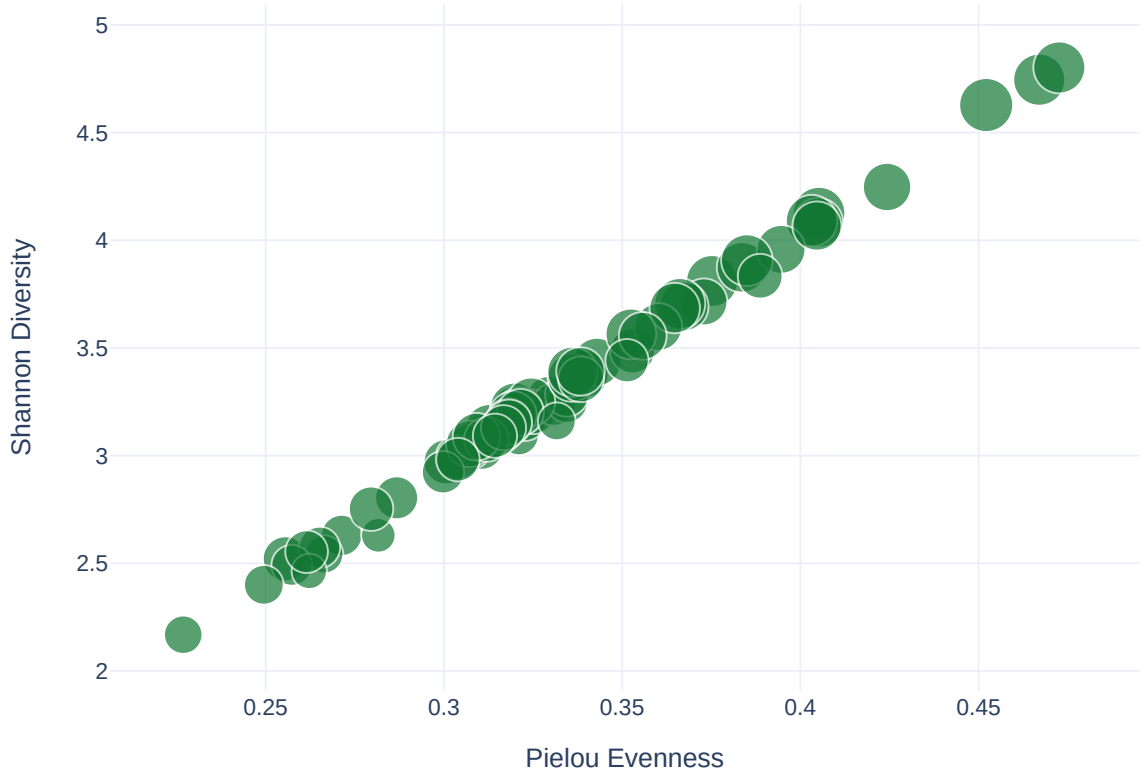


Fig. 3. Figure included as part of the EURYALE's MicroView report. It shows a scatterplot of Pielou Evenness and Shannon's diversity in each sample, represented by a dot. The size of each dot corresponds to the number of distinct species classified in each sample. In the HTML report the figure is interactive.

3 DISCUSSÃO

Ao averiguarmos o estado da arte dos fluxos de trabalho computacional presentes na área da metagenômica, pudemos concluir que a adoção de tecnologias de conteinerização e orquestradores de execução ainda é um tanto limitada. Essas limitações tornam-se mais evidentes sobretudo no contexto de metodologias de metagenômica livres de montagem. Tais metodologias são mais escassas e menos computacionalmente intensas quando comparadas às abordagens baseadas em montagem. Ainda assim, elas necessitam processamento adequado, de forma reprodutível, replicável e automatizável, especialmente por possibilitarem uma análise realizável em infraestruturas de menor porte e por possuírem melhor sensibilidade ao tratar dados com baixa cobertura (AYLING; CLARK; LEGGETT, 2020), ambos fatores comuns quando se tratando da produção e processamento de dados metagenômicos em ambientes com financiamento científico limitado, como no sul global.

No contexto de possíveis metodologias, decidimos então aplicar os princípios postos para a metodologia MEDUSA (MORAIS *et al.*, 2022), que apresentou resultados superiores a fluxos de trabalho semelhantes, além de ter obtido suas ferramentas a partir de curadoria manual, com rigorosos processos de benchmarking. Portanto, tomando vantagem da curadoria precedente, decidimos então re-implementar o MEDUSA, utilizando-se agora do gerenciador de pipelines Nextflow a de tecnologias de contêinerização Docker e Singularity.

Nextflow foi escolhido sobretudo devido à facilidade e rapidez de desenvolvimento, conjunta ao suporte multiplataforma mais abrangente que o orquestrador anteriormente escolhido para o MEDUSA, Snakemake. Ademais, Nextflow já foi selecionado como a melhor opção para desenvolvimento de fluxos de trabalho em comparações anteriores (JACKSON; KAVOUSSANAKIS; WALLACE, 2021) (CELEBI; MCDANIEL; REITER, 2023). Outro aspecto que levou à decisão por Nextflow foi a existência da comunidade nf-core, que realiza trabalhos de curadoria de fluxos de trabalho em bioinformática e fornecem templates para a criação de pipelines modulares, altamente parametrizáveis e que possibilitem melhor manutenção a longo prazo (EWELS, P. A. *et al.*, 2020).

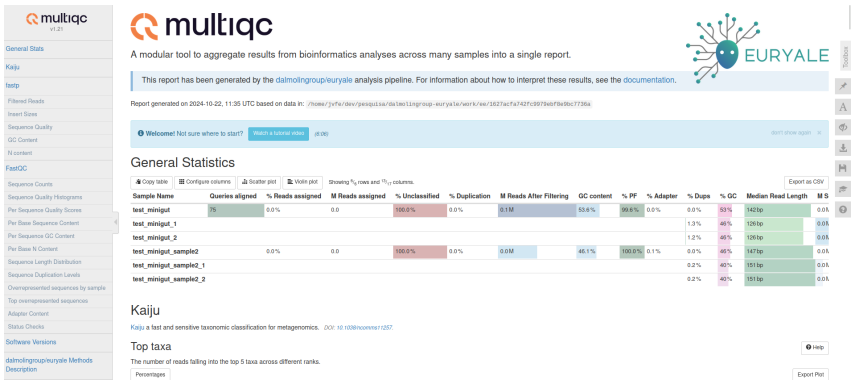
Apesar de reimplementarmos o MEDUSA por completo como EURYALE, também avançamos em algumas limitações do software original, sobretudo interpretabilidade dos dados e parametrização.

Quanto ao primeiro ponto, notávamos uma clara ausência, no MEDUSA, de visualizações e relatórios que forneçam informações preliminares e análises exploratórias acerca do conjunto de dados processado. No entanto, acessibilidade de dados através de relatórios e figuras interativas é essencial para tornar análises mais compreensíveis e reprodutíveis, sobretudo em um campo com complexidade alta de informação, com

muitos arquivos gerados por análise, como a metagenômica.

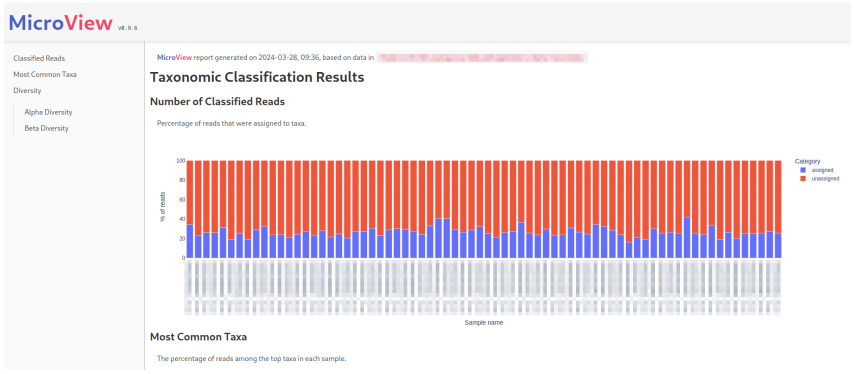
Nesse sentido, primeiro interligamos as diversas ferramentas agora presentes no EURYALE através do MultiQC, um agregador de informação de dados de bioinformática (EWELS, P. *et al.*, 2016) que permitiu disponibilizar um relatório interativo e configurável com informações a respeito das etapas de controle de qualidade, classificação taxonômica e do alinhamento que precede a anotação funcional (Figura 1).

Figura 1 – Exemplo do cabeçalho de um relatório gerado através da ferramenta MultiQC, implementada como parte integrante do EURYALE.



Ademais, através da implementação de uma nova ferramenta, MicroView, também inclusa no EURYALE, disponibilizamos também ao usuário métricas pré-calculadas de diversidade, ilustrando, de forma inicial e exploratória, o quadro geral resultante da classificação taxonômica (Figura 2).

Figura 2 – Exemplo do cabeçalho de um relatório gerado através da ferramenta Microview, escrita em Python e implementada como parte integrante do EURYALE.



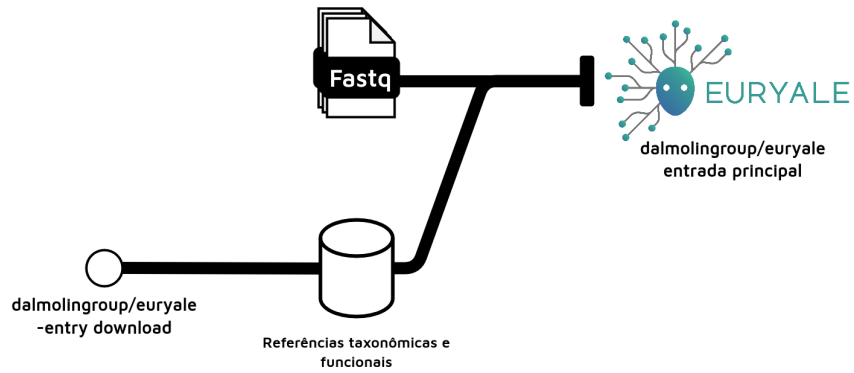
O MicroView também atua como um pacote na linguagem Python de programação, possibilitando sua execução fora do fluxo de trabalho. De forma geral, tais relatórios, com visualizações interativas, aumentam a acessibilidade de dados das análises realizadas com o EURYALE, facilitando não apenas a exploração inicial e a geração de descobertas científicas pelos pesquisadores usuários da metodologia,

mas também a exploração posterior dos resultados, promovendo re-análises (PERKEL, 2018).

Vale ressaltar também os aprimoramentos em padronização de software e parametrização trazidos da mudança do MEDUSA para o EURYALE. Ao implementarmos a metodologia, optamos por utilizar a infraestrutura da nf-core, uma comunidade que busca padronizar o desenvolvimento de fluxos de trabalho que utilizam Nextflow, com ênfase em modularidade, testes contínuos e documentação (EWELS, P. A. *et al.*, 2020). A comunidade implementa uma suíte de software em Python para a criação e manutenção de fluxos de trabalho (`https://nf-co.re/docs/nf-core-tools/installation`), que foi amplamente utilizada no desenvolvimento do EURYALE. Ao utilizarmos das ferramentas e diretrizes da nf-core, pudemos superar algo que foi a principal crítica de usuários em relação ao MEDUSA: A estaticidade das referências e a pouca parametrização.

Quanto ao primeiro ponto, optamos por uma solução que possibilite a utilização de bancos de dados referência equivalentes aos utilizados pelo MEDUSA, que é o que tomamos como padrão da análise. Nesse sentido, implementamos uma entrada alternativa ao fluxo de trabalho (`-entry download`‘), que possibilita que os usuários transfiram os bancos de dados referência para sua máquina de análise, alimentando assim o *pipeline* propriamente dito (Figura 3).

Figura 3 – Diagrama ilustrando como a entrada de download do EURYALE adquire os bancos de dados que alimentam a entrada principal. Com uma análise típica do zero sendo constituída por ambas etapas.

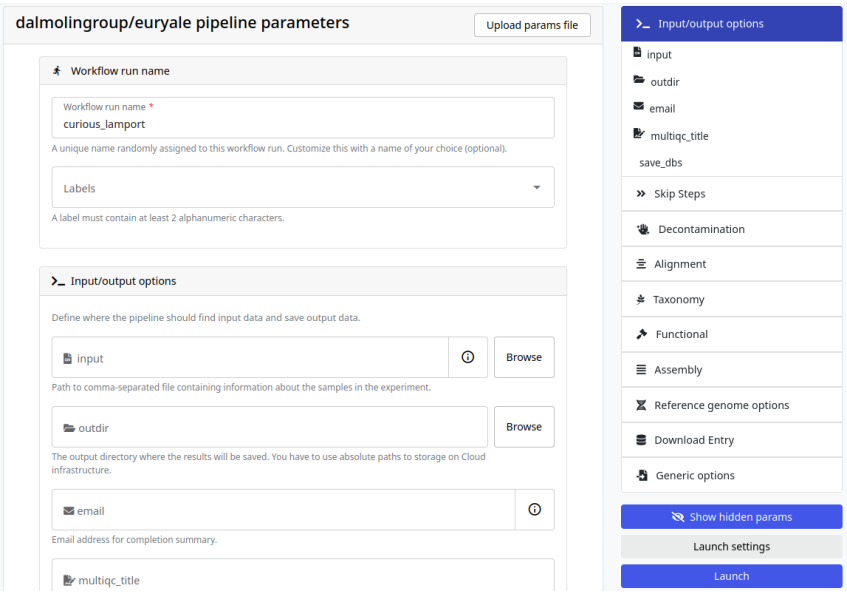


Apesar disso, devido à modularidade advinda do *template nf-core*, o EURYALE é modelado para funcionar com quaisquer outro banco de dado compatível com as suas ferramentas integrantes. Dessa maneira, o usuário pode tanto modificar os parâmetros de *download* na entrada específica, quanto modificar diretamente os parâmetros de referência presentes na entrada principal do *pipeline*, caso ele possua referências pré-adquiridas. Esses detalhes funcionais distinguem a usabilidade do EURYALE com a de seu precedente, não requerendo que os usuários alterem o código fonte para utilizar referências diferentes.

Além disso, adicionamos mais parâmetros ao EURYALE, parâmetros estes tanto que alterem as etapas de análise quanto aqueles que alteram os recursos computacionais alocados a estas etapas. A exemplo dessa mudança, podemos ver como a metodologia agora apoia um passo opcional de montagem, semelhante ao descrito no artigo original do MEDUSA, no entanto não aplicado diretamente na sua versão orquestrável original. O novo fluxo de trabalho também oferece mais opções de classificação taxonômica, além de parâmetros que possibilitam desabilitar diferentes passos de análise, como a descontaminação de leituras do hospedeiro, possibilitando que a abordagem se adapte a diferentes contextos de análise. Adicionalmente, outro ponto originado pela utilização do template *nf-core* são os diferentes rótulos de alocação de recursos, que permitem um gerenciamento fácil, da quantidade de recursos que pode ser alocada a cada passo de análise, através de um único arquivo de configuração. Essa adição permite que a metodologia seja executada em diferentes infraestruturas computacionais.

Em última instância, vale ressaltar como a utilização do *template nf-core* também possibilita a implantação do fluxo de trabalho na *Seqera Platform*, com a geração automática de uma interface gráfica, permitindo a usuários executar a metodologia sem necessariamente utilizarem uma interface em linha de comando, dado que, é claro, possuam cadastro na plataforma (Figura 4).

Figura 4 – Porção inicial da interface gráfica do EURYALE na Seqera Platform (<https://cloud.seqera.io/>). Através desse formulário usuários podem selecionar os parâmetros a serem utilizados para a execução do fluxo de trabalho.



4 CONCLUSÃO

Ressaltar como avançamos nas metodologias de meta com orquestração e conteinerização, priorizando princípios de software sustentável. Na mesma medida, nos baseamos em uma metodologia sólida já existente, aprimorando nesses conceitos além de possibilitar melhor interpretabilidade com a implementação de uma ferramenta adicional de reporting. No entanto, faltar como metodologias de integração de dados metagenomica/metatranscriptomica ainda são relativamente escassas, sobretudo considerando-se fluxos de trabalho automatizados e que sigam esses mesmos princípios.

REFERÊNCIAS

ALTSCHUL, Stephen *et al.* The anatomy of successful computational biology software. **Nature Biotechnology**, v. 31, n. 10, p. 894–897, out. 2013. Publisher: Nature Publishing Group. DOI: 10.1038/nbt.2721. Disponível em: https://www.nature.com/articles/nbt.2721.

AYLING, Martin; CLARK, Matthew D; LEGGETT, Richard M. New approaches for metagenome assembly with short reads. **Briefings in Bioinformatics**, v. 21, n. 2, p. 584–594, 23 mar. 2020. DOI: 10.1093/bib/bbz020. Disponível em: https://doi.org/10.1093/bib/bbz020.

BREITWIESER, Florian P.; LU, Jennifer; SALZBERG, Steven L. A review of methods and databases for metagenomic classification and assembly. **Briefings in Bioinformatics**, v. 20, n. 4, p. 1125–1136, 19 jul. 2019. PMID: 29028872 PMCID: PMC6781581. DOI: 10.1093/bib/bbx120.

CELEBI, Feridun Mert; MCDANIEL, Elizabeth; REITER, Taylor. Creating reproducible workflows for complex computational pipelines. **Arcadia Science**, 7 mar. 2023. Publisher: Arcadia Science. DOI: 10.57844/arcadia-cc5j-a519. Disponível em: https://research.arcadiascience.com/pub/perspective-reproducible-workflows/release/2.

CHIARELLO, Marlène *et al.* Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. **PLOS ONE**, v. 17, n. 2, e0264443, inverno 2022. Publisher: Public Library of Science. DOI: 10.1371/journal.pone.0264443. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0264443.

COMIN, Matteo *et al.* Comparison of microbiome samples: methods and computational challenges. **Briefings in Bioinformatics**, v. 22, n. 1, p. 88–95, 18 jan. 2021. DOI: 10.1093/bib/bbaa121. Disponível em: https://academic.oup.com/bib/article/22/1/88/5861761.

DI TOMMASO, Paolo *et al.* Nextflow enables reproducible computational workflows. **Nature Biotechnology**, v. 35, n. 4, p. 316–319, abr. 2017. Number: 4 Publisher: Nature Publishing Group. DOI: 10.1038/nbt.3820. Disponível em: https://www.nature.com/articles/nbt.3820.

DOUGLAS, Gavin M. *et al.* PICRUSt2 for prediction of metagenome functions. **Nature Biotechnology**, v. 38, n. 6, p. 685–688, jun. 2020. Publisher: Nature Publishing Group. DOI: 10.1038/s41587-020-0548-6. Disponível em: https://www.nature.com/articles/s41587-020-0548-6.

EWELS, Philip *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, v. 32, n. 19, p. 3047–3048, 1 out. 2016. DOI: 10.1093/bioinformatics/btw354. Disponível em: https://doi.org/10.1093/bioinformatics/btw354.

EWELS, Philip A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. **Nature Biotechnology**, v. 38, n. 3, p. 276–278, mar. 2020. Number: 3 Publisher: Nature Publishing Group. DOI: 10.1038/s41587-020-0439-x. Disponível em: https://www.nature.com/articles/s41587-020-0439-x.

JACKSON, Michael; KAVOUSSANAKIS, Kostas; WALLACE, Edward W. J. Using prototyping to choose a bioinformatics workflow management system. **PLOS Computational Biology**, v. 17, n. 2, e1008622, **springN** 2021. Publisher: Public Library of Science. DOI: 10.1371/journal.pcbi.1008622. Disponível em: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008622.

KADRI, Sabah *et al.* Containers in Bioinformatics: Applications, Practical Considerations, and Best Practices in Molecular Pathology. **The Journal of Molecular Diagnostics**, v. 24, n. 5, p. 442–454, 1 mai. 2022. DOI: 10.1016/j.jmoldx.2022.01.006. Disponível em: https://www.sciencedirect.com/science/article/pii/S1525157822000381.

KRAKAU, Sabrina *et al.* nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. **NAR Genomics and Bioinformatics**, v. 4, n. 1, lqac007, 1 mar. 2022. DOI: 10.1093/nargab/lqac007. Disponível em: https://doi.org/10.1093/nargab/lqac007.

LIU, Yong-Xin *et al.* A practical guide to amplicon and metagenomic analysis of microbiome data. **ProteinCell**, v. 12, n. 5, p. 315–330, 1 mai. 2021. DOI: 10.1007/s13238-020-00724-8. Disponível em: https://doi.org/10.1007/s13238-020-00724-8.

MAGNABOSCO, Cara *et al.* Toward a Natural History of Microbial Life. **Annual Review of Earth and Planetary Sciences**, v. 52, Volume 52, 2024, p. 85–108, 23 jul. 2024.

Publisher: Annual Reviews. DOI: 10.1146/annurev-earth-031621-070542. Disponível em: https://www.annualreviews.org/content/journals/10.1146/annurev-earth-031621-070542.

MANGUL, Serghei; MARTIN, Lana S. *et al.* Improving the usability and archival stability of bioinformatics software. **Genome Biology**, v. 20, n. 1, p. 47, 27 fev. 2019. DOI: 10.1186/s13059-019-1649-8. Disponível em: https://doi.org/10.1186/s13059-019-1649-8.

MANGUL, Serghei; MOSQUEIRO, Thiago *et al.* Challenges and recommendations to improve the installability and archival stability of omics computational tools. **PLOS Biology**, v. 17, n. 6, e3000333, 20 jun. 2019. Publisher: Public Library of Science. DOI: 10.1371/journal.pbio.3000333. Disponível em: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000333.

MARCHESI, Julian R.; RAVEL, Jacques. The vocabulary of microbiome research: a proposal. **Microbiome**, v. 3, n. 1, p. 31, 30 jul. 2015. DOI: 10.1186/s40168-015-0094-5. Disponível em: https://doi.org/10.1186/s40168-015-0094-5.

MÖLDER, Felix *et al.* Sustainable data analysis with Snakemake. **F1000Research**, v. 10, p. 33, 19 abr. 2021. PMID: 34035898 PMCID: PMC8114187. DOI: 10.12688/f1000research.29032.2. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8114187/.

MORAIS, Diego A. A. *et al.* MEDUSA: A Pipeline for Sensitive Taxonomic Classification and Flexible Functional Annotation of Metagenomic Shotgun Sequences. **Frontiers in Genetics**, v. 13, 2022. Disponível em: https://www.frontiersin.org/articles/10.3389/fgene.2022.814437.

PERKEL, Jeffrey M. Data visualization tools drive interactivity and reproducibility in online publishing. **Nature**, v. 554, n. 7690, p. 133–134, 30 jan. 2018. Bandiera_abtest: a Cg_type: Toolbox Publisher: Nature Publishing Group Subject_term: Authorship, Peer review, Publishing. DOI: 10.1038/d41586-018-01322-9. Disponível em: https://www.nature.com/articles/d41586-018-01322-9.

SALAZAR, Vinícius W *et al.* Metaphor—A workflow for streamlined assembly and binning of metagenomes. **GigaScience**, v. 12, giad055, 1 jan. 2023. DOI:

10.1093/gigascience/giad055. Disponível em: https://doi.org/10.1093/gigascience/giad055.

STRAUB, Daniel *et al.* Interpretations of Environmental Microbial Community Studies Are Biased by the Selected 16S rRNA (Gene) Amplicon Sequencing Pipeline. **Frontiers in Microbiology**, v. 11, 23 out. 2020. Publisher: Frontiers. DOI: 10.3389/fmicb.2020.550420. Disponível em: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2020.550420/full.

TREMBLAY, Julien; SCHREIBER, Lars; GREER, Charles W. High-resolution shotgun metagenomics: the more data, the better? **Briefings in Bioinformatics**, v. 23, n. 6, bbac443, 1 nov. 2022. DOI: 10.1093/bib/bbac443. Disponível em: https://doi.org/10.1093/bib/bbac443.

WOOLEY, John C.; GODZIK, Adam; FRIEDBERG, Iddo. A Primer on Metagenomics. **PLOS Computational Biology**, v. 6, n. 2, e1000667, **summerN** 2010. Publisher: Public Library of Science. DOI: 10.1371/journal.pcbi.1000667. Disponível em: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000667.