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 ©2024 IEEE

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].

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

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

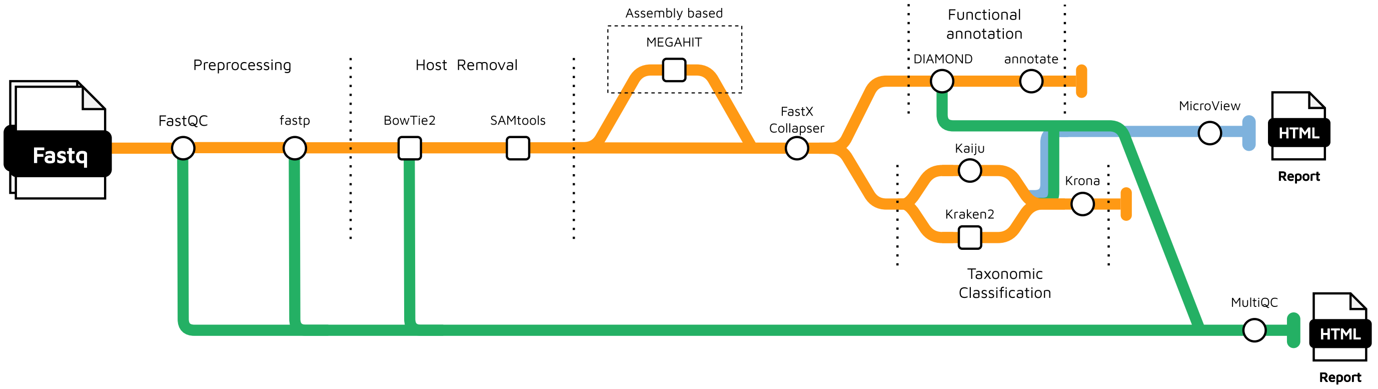

Fig. 1. Diagram showing each step of the EURYALE pipeline, starting with FastQ reads as input. Circles represent steps that are enabled by default, while squares represent steps that are optional or disabled by default.

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. 168 016, 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).

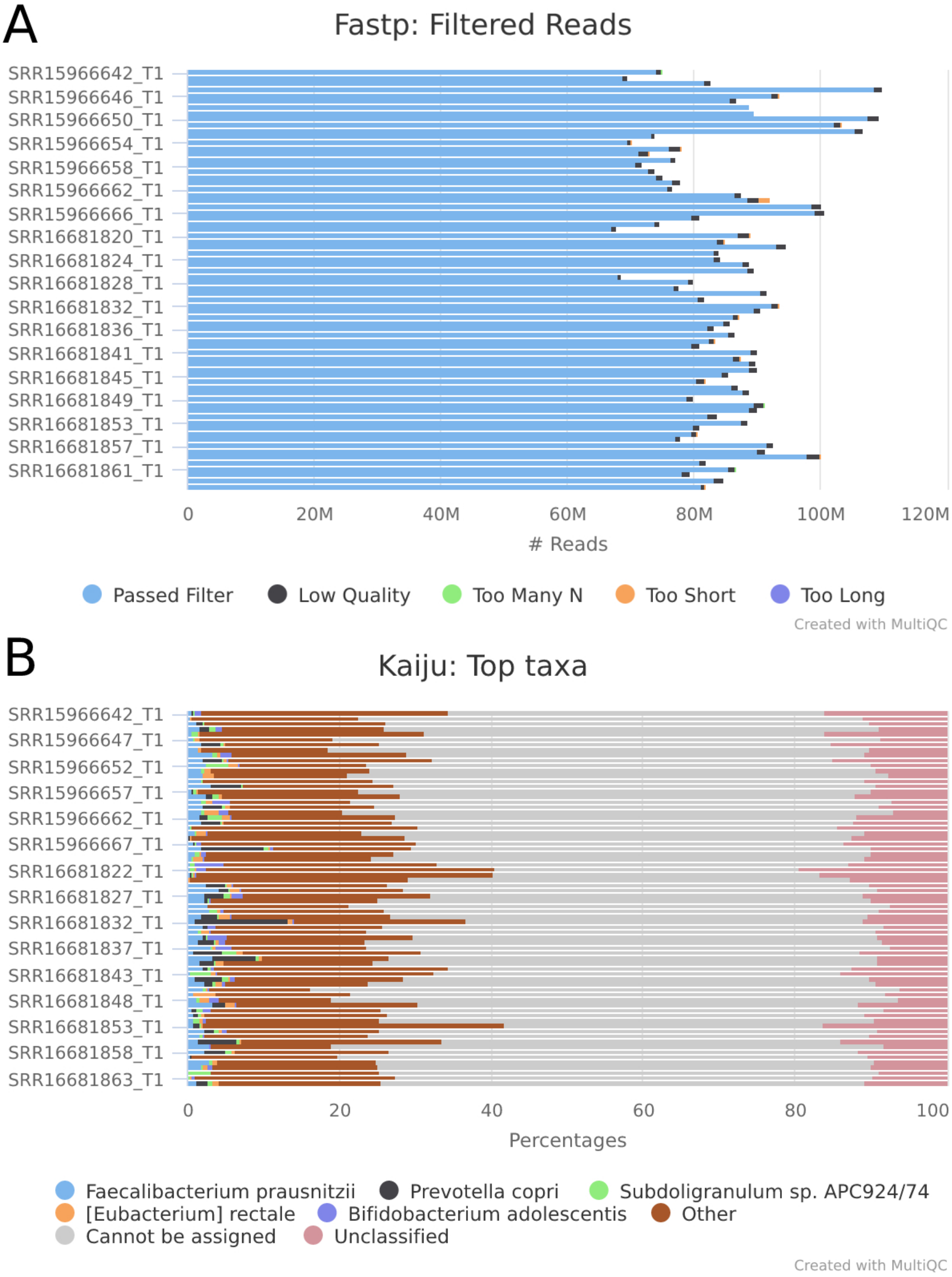

Fig. 2. Example figures included as part of EURYALE's MultiQC report. **A**: Figure for the quality control section, showing the number of reads filtered out in each sample by fastp and their respective thresholds for filtering. **B**: Figure for the taxonomic classification section, showing the percentage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

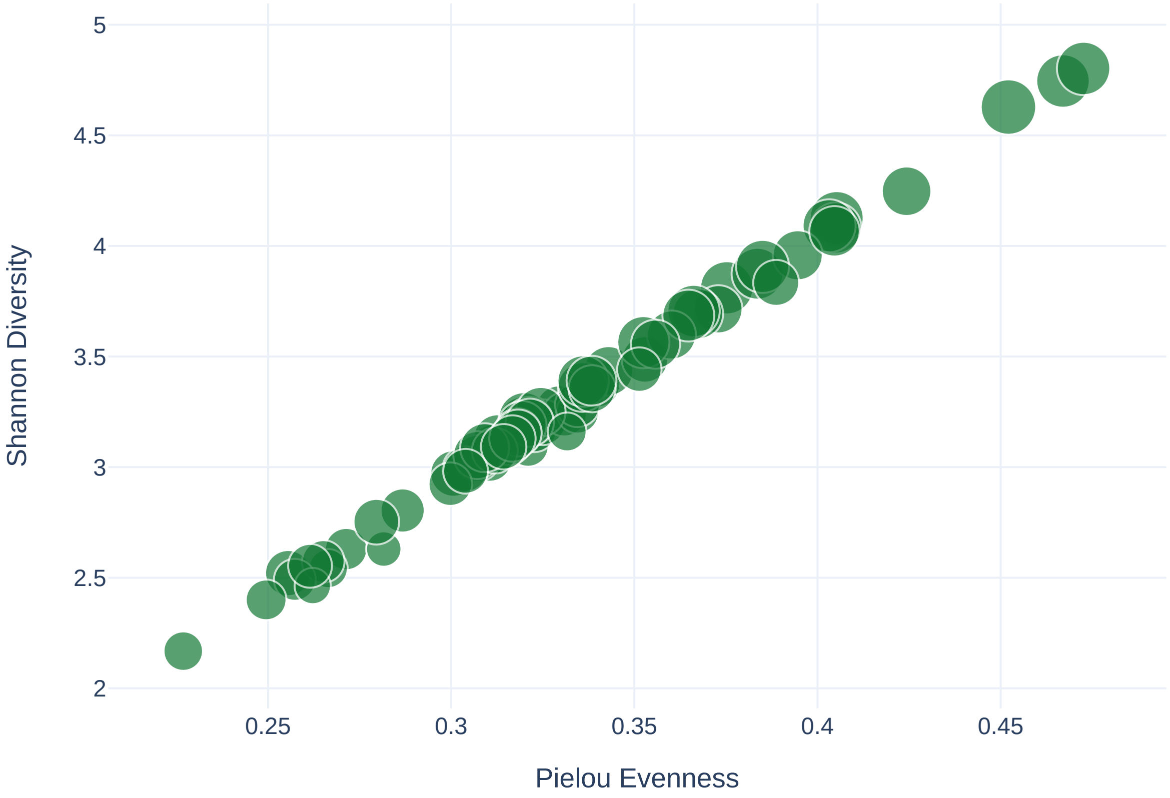

Fig. 3. Figure included as part of the EURYALE's MicroView report. It shows a scatterplot of Pielou Evenness and Shannon's diversity in each sample, represented by a dot. The size of each dot corresponds to the number of distinct species classified in each sample. In the HTML report the figure is interactive.