BioHackrXiv

BioHackEU25 report project 16: MiCoReCa (Microbiome Community Resource Catalogue) - Towards Centralized Curation and Integration of Microbiome Bioinformatics Resources

Vivek Ashokan [1], **Clara Emery** [2, 3], **Agnès Barnabé** [4, 5], **Valentin Loux** [4], **Christina Pavloudi** [8], **Paul Zierep** [9], **Nikolaos Strepis** [10], **and Bérénice Batut** [2, 11]

1 LABGeM (Laboratory of Bioinformatics Analyses for Genomics and Metabolism), Genoscope, IBFJ, DFR, CEA, 91000 Evry-Courcouronnes, France **2** IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France **3** ABiMS - Analysis and Bioinformatics for Marine Science Roscoff Marine Station **4** Migale bioinformatics facility, MaIAGE, BioInfomics, INRAE, 78350 Jouy-en-Josas, France **7** Ferments du Futur (US INRAE 1503), 91400 Orsay, France **8** European Marine Biological Resource Centre (EMBRC-ERIC), Paris, France **9** Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Koehler-Allee 106, D-79110 Freiburg, Germany **10** Department of Pathology and Clinical Bioinformatics, Erasmus MC Cancer Institute, Erasmus MC, Rotterdam, Netherlands **11** Plateforme AuBi, Mésocentre Clermont-Auvergne, Université Clermont Auvergne, Aubière, France

BioHackathon series:
BioHackathon Europe 2025
Berlin, Germany, 2025
Project 16

Submitted: 17 Dec 2025

Published by BioHackrXiv.org

Introduction

The rapid expansion of microbiome research has led to the development of countless bioinformatics tools, workflows, and databases. However, information about these resources remains scattered across disparate, often outdated catalogs, impeding their discovery and effective use (citation). To address this critical gap, the ELIXIR Microbiome Community (Finn et al., 2025) —a specialized group within ELIXIR, Europe's leading infrastructure for biological data —proposed the creation of **MiCoReCa (Microbiome Community Resource Catalogue)**. This open-access, dynamic catalog aims to centralize and streamline access to microbiome-related bioinformatics resources, including tools, workflows, training materials, and more.

MiCoReCa integrates resources from established platforms such as Bioconda (a package manager for bioinformatics software) (Grüning et al., 2018), bio.tools (the registry of bioinformatics tools and services developed by ELIXIR) (Ison et al., 2016, 2019) via the Research Software Ecosystem (RSEc) Atlas, WorkflowHub (a repository for scientific workflows, maintained by ELIXIR) (Gustafsson et al., 2025), and TeSS (the ELIXIR training materials database) (Beard et al., 2020). By aggregating resources from these sources, MiCoReCa ensures comprehensive coverage of the microbiome bioinformatics landscape.

This report presents the **work accomplished during the ELIXIR BioHackathon 2025**—an intensive collaborative event where researchers, developers, and bioinformatics experts come together to tackle scientific challenges. During this event, the MiCoReCa team initiated the development of the catalog, focusing on automating resource extraction (e.g., via weekly GitHub Actions scripts), filtering resources using community-defined keywords, and establishing a framework for collaborative curation, inspired by the Galaxy Codex (Zierep et al., 2024).

To maximize interoperability, MiCoReCa leverages **standardized ontologies** like EDAM (Ison et al., 2013), which provides a structured vocabulary for describing bioinformatics tools, workflows, and data. This ensures that resources in MiCoReCa are consistently annotated, making them easier to discover and integrate into research workflows.

A defining feature of MiCoReCa is its **community-driven curation process**, where experts collaboratively identify missing ontological terms and metadata, ensuring the catalog remains **accurate, up-to-date, and aligned with researchers' needs**.

Beyond serving as a vital resource for the microbiome field—enhancing research efficiency and reproducibility—MiCoReCa is designed as a **scalable and adaptable infrastructure**, potentially applicable to other ELIXIR Communities. This initiative underscores the ELIXIR Microbiome Community's commitment to **streamlining microbiome bioinformatics** and fostering collaboration across disciplines.

Objectives before the ELIXIR BioHackathon 2025

Before the ELIXIR BioHackathon 2025, the following objectives were defined to guide the development and implementation of MiCoReCa:

1. **Extract and Expose Microbiome Resources from the ELIXIR Ecosystem and Bioconda** to create a comprehensive inventory of microbiome resources
    1. **Identify Keywords for Resource Filtering** on platforms such as **RSEc-Atlas**, **WorkflowHub**, **TeSS**, and **Bioconda**.
    2. **Coordinate with the Research Software Ecosystem (RSEc)** to ensure the inclusion of the microbiome community in the **RSEc-Atlas**.
    3. **Extract and Filter Microbiome Resources**: Automate the extraction and filtering of microbiome resources by scrapping
        1. **RSEc-Atlas** and **BioConda** for tools,
        2. **WorkflowHub** for workflows,
        3. **TeSS** for training resources.
    4. **Identify Missing Microbiome Tools in bio.tools** by comparing resources in Bioconda and WorkflowHub to those listed in bio.tools.
    5. **Create a Microbiome Resource Catalog Page** to expose all extracted microbiome resources as a searchable and accessible catalog.
2. **Expand, Curate, and Improve Annotation of Microbiome Resources** to ensure the accuracy, relevance, and usability of the catalog
    1. **Curate Extracted Resources** by reviewing with Microbiome community the extracted and filtered resources to ensure their quality and relevance.
    2. **Improve Tool Annotations** on **bio.tools** using standardized ontologies such as **EDAM**.
    3. **Add Microbiome tools** found in **Bioconda** and **WorkflowHub** but missing in **bio.tools**.
    4. **Engage in discussions** to refine and expand **EDAM** terms related to **Topics**, **Formats**, **Operations**, and **Data** for microbiome analysis.
3. **Document the Process** —from resource extraction and filtering to curation and annotation— **for Reusability by Other Communities** to ensure the sustainability and scalability of MiCoReCa.

These objectives laid the foundation for the work accomplished during the ELIXIR BioHackathon 2025, ensuring a structured and collaborative approach to building MiCoReCa.

Achievements During ELIXIR BioHackathon 2025

The ELIXIR BioHackathon 2025 brought together a diverse group of participants, including both onsite and online contributors, fostering a collaborative environment to advance the MiCoReCa project. Coordination was streamlined through a dedicated **Slack channel**, where real-time discussions and updates took place, complemented by **daily morning meetings** to align on progress and priorities. To ensure clarity and efficiency, the project leads prepared a **coordination document** outlining task descriptions and a **tracking spreadsheet** to monitor

progress throughout the event. At the outset of the BioHackathon, a **GitHub repository** was established inside the RSEc GitHub organization to centralize code, documentation, and collaborative efforts, providing a structured platform for version control and teamwork. This framework enabled the team to effectively tackle the objectives, resulting in significant progress toward building a comprehensive and curated microbiome resource catalog.

In this section, we highlight the key accomplishments made during the BioHackathon, demonstrating how the collective efforts of the MiCoReCa team advanced the development of the Microbiome Community Resource Catalogue.

Establishing Community-approved Keywords For Microbiome Resources Discovery

A central objective of MiCoReCa was to define a set of keywords that accurately represent the diverse landscape of microbiome computational tools and workflows. To ensure the catalog aligns with both **user needs** and **scientific relevance**, the project engaged the microbiome research community, harnessing their collective expertise to curate a keyword set that reflects real-world research practices. The development of these keywords was guided by the need for **platform compatibility**, ensuring seamless integration with the query structures of key repositories such as the **RSEc Atlas**, **WorkflowHub**, and **Bioconda**.

The keyword strategy was designed around **three distinct types of terms**: - **EDAM terms**, focusing on **topics** ("Metagenomics," "Metatranscriptomics", "Metabarcoding") and **operations** ("Read binning") to ensure alignment with standardized ontologies. - **General keywords**, structured as **regular expressions** (e.g., `metage.*`, `microbiom.*`) to flexibly capture variations in terminology and maximize resource discovery. - **Acronyms** (**ITS**, **OTU**, **ASV**), which are widely used in microbiome research and essential for identifying relevant tools and workflows.

To maintain **domain specificity**, the keyword set was deliberately scoped to **microbiome research**, explicitly excluding terms related to **general microbiology** or **single-genome analysis**. By combining **community-driven input**, **technical precision**, and **domain expertise**, this keyword framework ensures that MiCoReCa serves as a **targeted, comprehensive, and user-centric resource** for the microbiome research community. The **full list of keywords** is available in the MiCoReCa GitHub repository, providing transparency and enabling further community contributions and refinements.

Automated Scraping, Filtering, and Community Curation: A Reproducible Pipeline for Microbiome Resource Discovery

To systematically identify, filter, and curate microbiome-specific resources within the ELIXIR ecosystem, we designed and implemented a **multi-stage, automated pipeline** (Figure 1). This approach ensures that only the most relevant tools and workflows are included in the MiCoReCa catalog, combining efficiency with scientific precision. The pipeline begins with an **automated scraping stage**, where scripts aggregate comprehensive lists of tools and workflows from key repositories such as **Bioconda** and **WorkflowHub**.

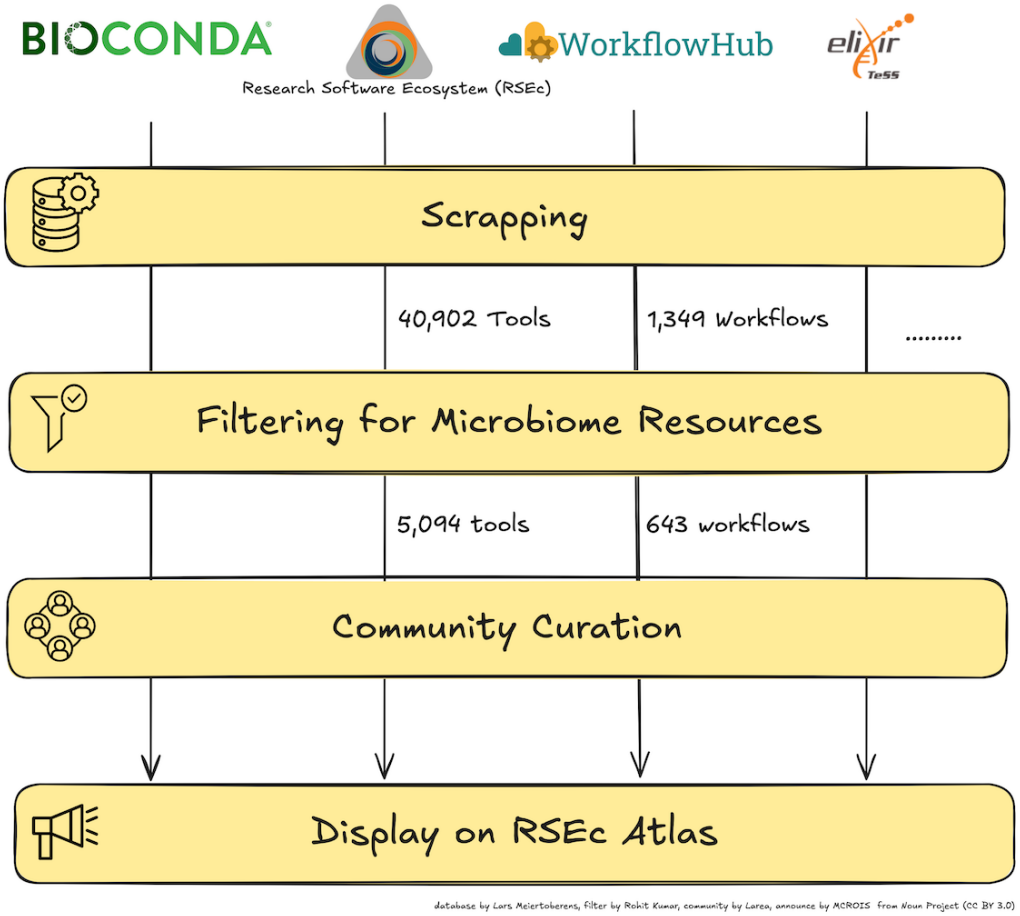

Figure 1: Workflow Diagram: The process for populating the catalog on RSEc Atlas, starting with scraping resources from Bioconda, RSEc, WorkflowHub, and TeSS, followed by filtering for microbiome-specific resources, community curation, and final display on the RSEc Atlas.

The next phase involves **automated filtering**, where a rule-based decision-tree logic (Figure 2) systematically evaluates metadata fields—including **EDAM topics, tags, keywords, and free-text descriptions**—against a predefined set of microbiome-specific keywords. This step refines the dataset of **tools** and **workflows**, ensuring only microbiome-relevant resources advance to the final stage. Resources that pass this automated screening are then subjected to **community curation**, where domain experts review and validate them before they are published on the **RSEc Atlas**.

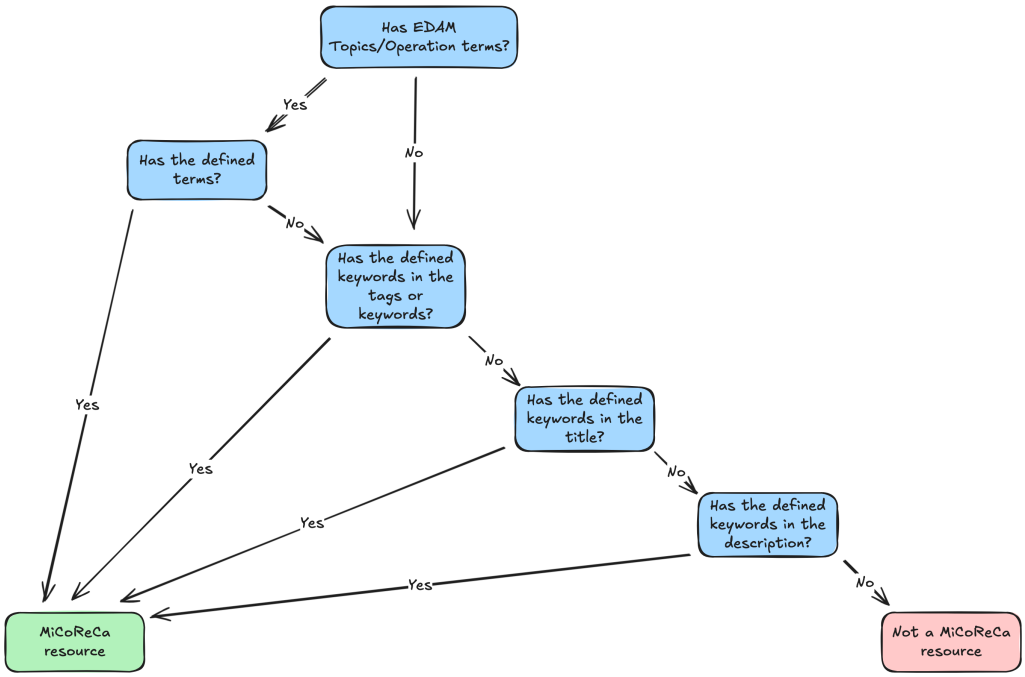

Figure 2: Filtering Logic: A flowchart illustrating the decision-tree process for classifying resources. The system checks for EDAM topics, keywords in tags, titles, and descriptions. A match at any step classifies the resource as a MiCoReCa resource.

To ensure the pipeline remains **up-to-date and reproducible**, we implemented **weekly automated scraping and filtering** via **GitHub Actions**, inspired by the **Galaxy Codex** model. Automated notifications are sent to the community for curation, ensuring continuous expert input, while **comprehensive documentation** of the entire process enables adoption by other ELIXIR communities and initiatives. This pipeline not only streamlines the discovery of microbiome resources but also establishes a **scalable, community-driven framework** for maintaining and expanding the MiCoReCa catalog.

Manual Curation of Extracted Tools and Workflows for Microbiome Relevance

ollowing the automated scraping and keyword-based filtering of resources from **bio.tools**, **WorkflowHub**, **Bioconda**, and **TeSS**, a rigorous **manual curation process** was conducted to ensure the inclusion of only the most relevant and high-quality resources for the microbiome research community. The primary objective of this curation was to refine the dataset by excluding resources that, despite passing the initial keyword filters, were not directly applicable to microbiome analysis. For instance, resources focused on **single-genome analysis**, **whole-genome sequencing**, **single-end ChIP-Seq data**, or other non-microbiome-specific applications were systematically removed.

To maintain a **focused and scientifically robust catalog**, resources were retained only if they aligned with one or more of the **key steps in microbiome data analysis**, including: (i) Quality Control (QC) and Filtering, (ii) Denoising or Clustering, (iii) Taxonomic Assignment and Phylogenetic Tree Construction, (iv) Functional Annotation, (v) Data Normalization, (vi) Diversity Analysis, (vii) Comparative Statistics, (viii) Association and Predictive Modeling. The **detailed curation protocol** is fully documented in the project's **GitHub repository**, ensuring transparency and reproducibility for future contributions.

Bioconda Tool Recipe Curation

The initial dataset of Bioconda recipes consisted of **10,136 recipes**. This set was reduced to **36 potential microbiome Bioconda recipes**. Unfortunately, **manual curation of the recipes** could not be initiated during the event due to **limited human resources**. This remains a priority for future work to ensure comprehensive coverage of microbiome tools.

bio.tools Tools Curation

The **bio.tools** (via RSEc-Atlas) yielded **over 4,000 potential microbiome tools** after filtering. Due to the **sheer volume** and **time constraints**, a **community-driven curation process** was initiated, with contributions expected from the microbiome research community in the coming months. To date, **201** tools have been reviewed with **169** tools confirmed as directly relevant to microbiome research. Additionally, the **integration of missing tools** identified from other platforms (e.g., WorkflowHub) could not be completed within the BioHackathon timeline, as the process required more extensive validation and coordination.

WorkflowHub Workflow Curation

The initial dataset from **WorkflowHub** consisted of **1,349 workflows**, which was reduced to **295 candidates** after automated filtering. Given the manageable size of this subset, a **manual review** was conducted, resulting in **208 workflows** being curated to date. Of these, **122 workflows** were confirmed as directly relevant to microbiome research and included in the final catalog.

This **multi-step curation process** ensures that MiCoReCa provides a **high-quality, relevant, and community-vetted** collection of microbiome bioinformatics resources, while also identifying areas for future improvement and expansion.

Expanding EDAM Ontology to Support the Microbiome Community

An ongoing discussion is focused on **enhancing the EDAM ontology** to better reflect the evolving needs of the microbiome research community. As MiCoReCa progresses, it has become evident that **additional terms**—particularly those related to **taxonomic classification** and **contig binning**—are essential for comprehensive resource annotation. These terms are proposed to be integrated as **topics** and/or **operations**, respectively, ensuring that the ontology accurately captures the nuances of microbiome data analysis.

The **expansion of EDAM** is anticipated to be an iterative process, with further **topics and operations** likely to emerge as MiCoReCa continues to grow and incorporate feedback from its users. To advance this effort, **collaborative discussions** with key members of the **EDAM consortium** and with the ELIXIR Microbiome will be prioritized in the coming months. This initiative aims to establish a **standardized, community-driven vocabulary** that enhances resource discovery, interoperability, and reproducibility in microbiome research.

Further Steps After the BioHackathon

The ELIXIR BioHackathon 2025 marked a significant milestone in the development of MiCoReCa, but the project's evolution is far from complete. Moving forward, our focus will be on **finalizing technical implementations**, **deepening community engagement**, and **expanding the catalog's reach** to other scientific domains. These efforts will ensure that MiCoReCa remains a **dynamic, high-quality, and sustainable resource** for microbiome research and beyond.

Finalizing the Implementation

To ensure the **long-term sustainability and functionality** of MiCoReCa, we will prioritize several key technical enhancements. First, we will **integrate support for TeSS**, enabling the inclusion of **training materials** alongside tools and workflows, thereby providing users with a more comprehensive resource hub. Another critical step will be **finalizing the data integration process with the RSEc Atlas**, which involves establishing a **semi-automated pipeline using GitHub Actions** to seamlessly update and display curated resources. This process will be developed in close collaboration with the RSEc Atlas team to incorporate **community-specific features** that enhance usability.

We also plan to introduce an **AI-assisted evaluation step**, leveraging **Large Language Models (LLMs)** to improve metadata annotation and resource classification. This will not only enhance the accuracy of resource descriptions but also streamline the curation process. Additionally, we will **expand the catalog by incorporating resources from the Open and Sustainable AI (OSAI) Ecosystem**, ensuring that MiCoReCa includes cutting-edge AI-driven tools relevant to microbiome research. Finally, we will **refactor and strengthen the codebase** by implementing **unit tests** and improving the overall structure, ensuring robustness and ease of maintenance for future development.

Collaboration with the ELIXIR Microbiome Community

Engagement with the **ELIXIR Microbiome Community** will remain central to MiCoReCa's growth. A key priority will be the **ongoing curation of resources and their metadata**, where community experts will review, validate, and enrich the information associated with each tool and workflow. This collaborative effort will ensure that the catalog remains **accurate, relevant, and aligned with the needs of researchers**.

Another important task will be **identifying and adding missing microbiome tools to bio.tools**, thereby improving the visibility and accessibility of these resources within the broader bioinformatics community. Additionally, we will continue our **collaboration with the EDAM consortium** to refine and finalize the **new EDAM terms** proposed during the BioHackathon. This will ensure that the ontology evolves to better represent the nuances of microbiome research, supporting more precise resource annotation and discovery.

Expanding Beyond the ELIXIR Microbiome Community

The **modular and well-documented nature** of MiCoReCa's pipeline presents an opportunity to extend its impact beyond the microbiome community. We will work with **Single-Cell Omics and Biodiversity communities** to adapt and generalize the MiCoReCa framework, making it applicable to other biological research domains. This effort will not only broaden the utility of the catalog but also foster **cross-disciplinary collaboration** and resource sharing.

Furthermore, components of the MiCoReCa codebase will be **directly integrated into the RSEc Atlas**, enabling other ELIXIR communities to adopt and build upon this framework. By doing so, we aim to create a **versatile, cross-domain resource catalog** that supports a wide range of scientific research endeavors.

Through these efforts, MiCoReCa will continue to evolve as a **comprehensive, sustainable, and community-driven resource**, setting a standard for collaborative bioinformatics tool curation and serving as a model for other scientific communities.

Conclusion and Perspectives

The **ELIXIR BioHackathon 2025** successfully established **MiCoReCa (Microbiome Community Resource Catalogue)**, a structured pipeline for curating microbiome-specific bioinformatics tools, workflows, and training materials from the ELIXIR ecosystem. A key achievement

was the development of a **community-driven keyword**, combining **EDAM ontology terms**, **regular expressions**, and **domain-specific acronyms** to ensure both **precision** and **comprehensive coverage** of microbiome research needs. This approach demonstrated the value of **collaborative expertise** in creating a resource tailored to real-world scientific workflows.

However, the project also highlighted **critical challenges** in scaling manual curation. While automated filtering effectively narrowed down resources, **manual validation** was only feasible for smaller datasets, such as **WorkflowHub**, where **62 high-quality workflows** were curated from an initial pool of 1,349. Larger repositories like **bio.tools** (over 4,000 resources) and **Bioconda** remained **partially or entirely uncurated** due to **time and resource constraints**, emphasizing the need for a **sustained, community-wide effort** to achieve comprehensive coverage.

Moving forward, our focus will be on **deepening community engagement** to complete the curation of remaining resources and **strengthening the infrastructure** by integrating new **microbiome-specific EDAM terms** and exploring **AI-assisted curation tools** to streamline validation. The **MiCoReCa pipeline**, thoroughly documented and modular, serves as a **reusable framework** that can be adapted by other **ELIXIR communities**—such as those in **single-cell omics** or **biodiversity**—to enhance resource discoverability and interoperability.

Ultimately, MiCoReCa not only addresses a critical gap in microbiome research but also sets a **scalable precedent** for collaborative bioinformatics resource curation. Its success will depend on **ongoing community contributions** and **technological advancements**, ensuring it evolves into a **cornerstone resource** for microbiome science and beyond. We encourage researchers and bioinformatics experts to **engage, adopt, and expand** this initiative, fostering a more **connected and efficient** bioinformatics ecosystem.

Acknowledgements

This work was developed as part of BioHackathon Europe 2025. This work was supported by ELIXIR, the research infrastructure for life science data. The French Institute of Bioinformatics (IFB) was founded by the Future Investment Program subsidized by the National Research Agency, number ANR-11-INBS-0013. This work received state aid managed by the National Research Agency under France 2030 for structural research equipment / EQUIPEX+ with reference ANR-21-ESRE-0048.

References

Beard, N., Bacall, F., Nenadic, A., Thurston, M., Goble, C. A., Sansone, S.-A., & Attwood, T. K. (2020). TeSS: A platform for discovering life-science training opportunities. *Bioinformatics*, *36*(10), 3290–3291. **[cito:citesAsAuthority]**

Finn, R. D., Balech, B., Burgin, J., Chua, P., Corre, E., Cox, C. J., Donati, C., Dos Santos, V. M., Fosso, B., Hancock, J.others. (2025). Establishing the ELIXIR microbiome community. *F1000Research*, *13*, ELIXIR–50. **[cito:citesForInformation]**

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., & Team, B. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. **[cito:citesAsAuthority]**

Gustafsson, O. J. R., Wilkinson, S. R., Bacall, F., Soiland-Reyes, S., Leo, S., Pireddu, L., Owen, S., Juty, N., Fernández, J. M., Brown, T.others. (2025). WorkflowHub: A registry for computational workflows. *Scientific Data*, *12*(1), 837. **[cito:citesAsAuthority]**

Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Ménager, H., Kalaš, M., Schwämmle, V., Grüning, B., Beard, N., Lopez, R.others. (2019). The bio. Tools registry of software tools and data resources for the life sciences. *Genome Biology*, *20*(1), 164. **[cito:citesAsAuthority]**

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., & Rice, P. (2013). EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, *29*(10), 1325–1332. **[cito:citesAsAuthority]**

Ison, J., Rapacki, K., Ménager, H., Kalaš, M., Rydza, E., Chmura, P., Anthon, C., Beard, N., Berka, K., Bolser, D.others. (2016). Tools and data services registry: A community effort to document bioinformatics resources. *Nucleic Acids Research*, *44*(D1), D38–D47. **[cito:citesAsAuthority]**

Zierep, P., Batut, B., Kalaš, M., Kayikcioglu, T., Nasr, E., Soranzo, N., Thang, W. C., Wang, J., & Gustafsson, J. (2024). *How to increase the findability, visibility, and impact of Galaxy tools for your scientific community*. https://doi.org/10.37044/osf.io/qjbxc **[cito:citesAsAuthority]**

Author contributions

Vivek Ashokan: Colead, Conceptualization, Development, Data curation Clara Emery: Writing – original draft, Development, Data curation Agnès Barnabé: Development, Manual curation Valentin Loux: Manual curation Christina Pavloudi: Manual curation Paul Zierep: Development Nikolaos Strepis: Colead, Conceptualization, Manual curationWriting – review & editing Bérénice Batut: Colead, Conceptualization, Development, Manual curation, Writing – review & editing