# BITS :: Call for Abstracts 2024 - Oral communication

| | |
|---|---|
| *Type* | Oral communication |
| *Session* | Bioinformatics Core Facilities and Research Infrastructures |
| *Title* | Driving Microbial Research Forward: Scalable Bioinformatics Workflows as a Service in the Italian Microbial Resource Research Infrastructure |
| *All Authors* | Sandro Gepiro Contaldo (1), Francesco Venice (2), Lorenzo Bosio (1), Iacopo Colonelli (1), Raffaele Adolfo Calogero (3), Ivan Molineris (2), Ilario Ferrocino (4), Ugo Ala (5), Simone Donetti (1), Antonio D'Acierno (6), Giovanna Cristina Varese (2), Marco Beccuti (1). |

*Affiliation*

(1) Department of Computer Science, Università degli Studi di Torino, Turin, Italy.
(2) Department of Life Sciences and Systems Biology, Università degli Studi di Torino, Turin, Italy
(3) Department of Molecular Biotechnology and Health Sciences, Università degli Studi di Torino, Turin, Italy
(4) Department of Agricultural, Forest and Food Sciences, Università degli Studi di Torino, Turin, Italy
(5) Department of Veterinary Sciences, Università degli Studi di Torino, Turin, Italy
(6) Institute of Food Sciences, CNR-ISA, Avellino, Italy

*Motivation*

The Italian Node of the Microbial Resource Research Infrastructure (MIRRI-IT), overseen by the Joint Research Unit MIRRI-IT, plays a pivotal role in advancing microbial research excellence in Italy. Serving as the central hub for disseminating information, data, and services related to Italian microbial collections, MIRRI-IT fosters collaboration and facilitates knowledge exchange among various stakeholders.
In 2022, the Italian government acknowledged the strategic significance of MIRRI.IT, providing €17M in funding from the NextGeneration EU-funded PNRR for the SUS-MIRRI.IT project. This initiative aims to fortify MIRRI-IT by establishing a unified platform for accessing Italian Microbial Biological Resource Collections and their associated services, ensuring its long-term sustainability. Microbial resources are associated with large amounts of metadata to be managed, and often with high-throughput sequencing datasets aimed at valorizing their genetic potential. Thus, this work presents and discusses the implementation choices made to deploy scalable and portable bioinformatics workflows as online services within the project.

*Methods*

The implementations of SUS-MIRRI.IT bioinformatics workflows as services have required orchestrating various software and systems within an HPC-cloud convergence architecture (Fig.1A). In this setup, the Cloud system oversees the deployment of both the service front- and back-end, while the HPC system manages the heavy computational tasks of the service.
In detail, to ensure the accessibility and usability of the provided service, we designed and implemented a tailored web user-interface using Next.js (Fig.1B).
This framework is well-known for its robust capabilities in constructing high-quality server-side statically generated web applications. It improves code management and boosts reusability by streamlining page rendering with lazy loading and automatic code-splitting.
Differently, the workflow implementation uses the StreamFlow [1] framework, a Workflow Management System written in Python and based on the Common Workflow Language standard. StreamFlow's capability to enable concurrent execution of multiple communicating tasks in a multi-agent ecosystem allowed us to adopt a micro-service approach, encapsulating each task within a Docker image. Subsequently, the execution of these workflows is managed within the HPC system using SLURM [2], renowned for its highly scalable job-scheduling capabilities that enable the simultaneous execution of multiple workflows.
According to this the workflow execution always begins with the user uploading the input data and parameters through a web interface. Subsequently, the analysis is scheduled and executed once computational resources become available on the HPC. Upon completion, users receive a notification prompting them to download the resulting output.

*Results*

To show the effectiveness of the proposed approach, we present an example workflow designed for microbial genome assembly demonstrating its scalability by varying the number of available computational resources. In detail, this workflow supports both long- and short-reads assembly and harnesses various assembler tools (such as Canu [3], Flye [4], and wtdbg2 [5]) to improve the outcome quality. Its completed schema is reported in Fig.1C.
The experiments took as input data 16 yeast libraries, each containing an average of 400k reads generated with Oxford Nanopore Technology. The analysis was carried out by incrementally increasing the number of HPC nodes, where each node is equipped with 2 CPUs Intel(R) Xeon(R) E5-2697 2.3GHz - 18 cores and 125 GB RAM.
Fig.1D shows the achieved speed-up as the number of nodes varies. Specifically, the first bar represents the execution time using a single node, while the subsequent bars represent the execution times using 15 and 30 nodes, respectively.
These results indicate promising performance, with a speed-up of ~10x achieved when utilizing 15 nodes, while ~12x with 30 nodes.

*Info*

[1] I. Colonnelli, B. Cantalupo, I. Merelli and M. Aldinucci, "StreamFlow: Cross-Breeding Cloud

With HPC," in IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 4, pp. 1723-1737, 1 Oct.-Dec. 2021, doi: 10.1109/TETC.2020.3019202.
[2] Slurm: Simple Linux Utility for Resource Management, A. Yoo, M. Jette, and M. Grondona, Job Scheduling Strategies for Parallel Processing, volume 2862 of Lecture Notes in Computer Science, pages 44-60, Springer-Verlag, 2003.
[3] Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. (2017). doi:10.1101/gr.215087.116
[4] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel Pevzner, "Assembly of Long Error-Prone Reads Using Repeat Graphs", Nature Biotechnology, 2019 doi:10.1038/s41587-019-0072-8
[5] Ruan, J. and Li, H. (2019) Fast and accurate long-read assembly with wtdbg2. Nat Methodsdoi:10.1038/s41592-019-0669-3

| | |
|---|---|
| *filename* | Workflow_BITS.jpg |
| *Figure* | |


| | |
|---|---|
| *Availability* | https://github.com/qBioTurin/WorkflowAssembly |

## Dissemination Material

*Social*

@QbioGroup

*Summary*

Power up your microbial research with Scalable Bioinformatics Workflows as a Service! Now, anyone can tap into high computational resources, no tech expertise needed. Let's turbocharge discovery together!
#Bioinformatics #ResearchRevolution

## Corresponding Author

| | |
|---|---|
| *Name, Surname* | Sandro Gepiro, Contaldo |
| *Email* | sandrogepiro.contaldo@unito.it |

| | |
|---|---|
| *Submitted on* | 02.05.2024 |