

Disclaimer:

This deliverable has been submitted to the European Commission and is currently under review. The final version after the approval may differ.



D3.3 – Data infrastructure for metagenomic analysis

Project Acronym:	MOBILES
Project Title:	Monitoring and Detection of Biotic and Abiotic Pollutants by Electronic Plants and Microorganisms Based Sensors
Grant Agreement:	101135402
Project coordinator:	prof. Evangelos Hristoforou dr. Angelo Ferraro National Technical University of Athens (NTUA)
Call:	HORIZON-CL6-2023-ZEROPOLLUTION-01.
Topic:	HORIZON-CL6-2023-ZEROPOLLUTION-01-6
Type of action:	HORIZON Research and Innovation Actions
Granting Authority:	European Research Executive Agency

Project: 101135402 — Mobiles — HORIZON-CL6-2023-ZEROPOLLUTION-01

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be responsible for them.



**Funded by
the European Union**

Deliverable D3.3 – Data infrastructure for metagenomics analysis

Short summary:	This document describes the hardware and software architecture of the specialized data infrastructure for metagenomics analysis, the integration of software processing and reporting tools, and the automatization of the data analysis. The work presented is part of WP3 and in particular Task 3.4: <i>Data infrastructure</i> The document describes the optimization, parallelization, and sequencing mechanisms implemented by the data infrastructure platform. This includes handling of data transactions and commitments when employing multiple databases, data consistency assurance by means of a transaction broker in charge of guaranteeing the consistency and interoperability of the data handled within the pipeline. The platform provides results visualization and interaction with the obtained data, notifies researchers about the status of their analyses, indicating which phase of the studies they are in, and alerts them to any issues that may arise during the pipeline processing. One of the analytical pipelines this infrastructure is designed to support is MeToReS (Metatranscriptomics Toward Ecosystem ResponseS), a Snakemake-based pipeline for soil metatranscriptomics specifically developed by WP3 MOBILES partners.
Due date:	31/08/2026
WP, leader:	[WP 3, CNR-ISAFOM]
Authors:	[WP 3, UPNA]

Dissemination Level

PU Public	☒
SEN Sensitive	☐

Document history

Version	Date	Name	Chapters edited	Reason for change
V1.0	31/07/2026	José Javier Astrain	All	Document prepared / Original Version
V1.1	04/08/2026	Lorenzo Pin	All	Document reviewed
V1.2	11/08/2026	Giuseppe Puglia	All	Document reviewed
V1.3	12/08/2026	José Javier Astrain	All	Document reviewed
V2	14/08/2026	Angelo Ferraro	All	Document reviewed / approved



List of participants

Participant No	Participant organisation name	Country
1 Coordinator	ETHNICON METSOVION POLYTECHNION (NTUA)	GR
2	CONSIGLIO NAZIONALE DELLE RICERCHE (CNR)	IT
3	INSTITUT NATIONAL DE RECHERCHE POUR L'AGRICULTURE, L'ALIMENTATION ET L'ENVIRONNEMENT (INRAE)	FR
4	UNIVERSITA DEGLI STUDI DI ROMA LA SAPIENZA (UR)	IT
5	EDEN TECH (EDEN) <i>Terminated consortium membership effective June 2026</i>	FR
6	UNIVERSIDAD PUBLICA DE NAVARRA (UPNA)	ES
7	INSTYTUT UPRAWY NAWOZENIA I GLEBOZNAWSTWA, PANSTWOWY INSTYTUT BADAWCZY (ISSPC)	PL
8	THE AGRICULTURAL RESEARCH ORGANISATION OF ISRAEL - THE VOLCANI CENTRE (ARO)	IL
9	UNIVERSITE DE BORDEAUX (UBx)	FR
10	TECHNOLOGIKO PANEPISTIMIO KYPROU (CUT)	CY
11	HEMIJSKI FAKULTET, UNIVERZITET U BEOGRADU (UBE)	RS
12	MAT4NRG-GESELLSCHAFT FUR MATERIALIEN UND ENERGIEANWENDUNGEN MBH (Mat)	DE
13	TECHNISCHE UNIVERSITAT CLAUSTHAL (TUC)	DE
14	GRANT GARANT SRO (GG)	CZ
15	CENTRUM BADAN I INNOWACJI PRO-AKADEMIA STOWARZYSZENIE (RICPA)	PL





MOBILES

Content

1.	Summary.....	5
2.	MOBILES project short description.....	6
3.	Introduction: Genomic Databases.....	9
4.	MOBILES platform overview.....	11
4.1.	MeToReS Pipeline.....	11
5.	Hardware infrastructure	12
6.	Platform virtualization.....	15
6.1.	Docker vs Conda	17
6.2.	Conda's election	18
6.3.	Platform description.....	18
7.	Operational modes	21
7.1.	DevOps Access Mode.....	21
7.2.	Standard User Access Mode	22
8.	Conclusions	25
9.	List of Figures.....	26
10.	List of Tables.....	26
11.	List of Abbreviations	26
12.	Project Consortium.....	28





MOBILES

1. Summary

Metagenomic and metatranscriptomic sequencing generate very large and heterogeneous datasets whose biological interpretation depends not only on the analytical methods applied but also on the data and computational infrastructure that stores, organises and processes them. A single soil metatranscriptome can comprise tens of gigabytes of raw reads, and a study like MOBILES spanning many sites and conditions rapidly accumulates terabytes of raw data, intermediate files, reference databases and results. Turning this volume of data into reproducible biological insight requires a dedicated environment able to store the data reliably, run complex analytical workflows efficiently, and make both the data and the results accessible to researchers with different levels of technical expertise.

Central to this process are the databases on which metagenomic analysis relies. On one hand, the analyses depend on large public reference databases — taxonomic, functional and, in this project, antimicrobial- and metal/biocide-resistance databases — against which the sequenced genes are classified and annotated. On the other hand, each project must build and maintain its own repository that collects and stores the raw sequence data, the processed outputs and the associated sample metadata. Building and sustaining such databases is far from trivial: the data are voluminous and heterogeneous, they must be curated, versioned and described with consistent metadata, and they must remain accessible and reusable over the long term. Most demanding of all, the analyses that populate and query them must be reproducible across time, operators and computing platforms — yet reference databases are continuously updated, software tools evolve and carry conflicting dependencies, and computing resources are finite, so that guaranteeing that the same input yields the same result years later is a genuine engineering challenge.

These requirements cannot be met by analytical software alone; they call for a purpose-built data and computational infrastructure. Such an infrastructure must combine high-performance computing and accelerator resources, a reliable multi-tier storage system, virtualisation and reproducible software environments, an automated workflow-orchestration layer, and controlled, secure access for multiple users. The present deliverable describes the infrastructure designed and deployed within the MOBILES project to meet these needs.

This deliverable is the computational counterpart of the analytical work carried out in WP3 and integrates with other WP3's deliverables. Deliverable D3.1 defined the sampling campaign and the extraction of nucleic acids; deliverable D3.2 defined the sequencing protocols and the reference analytical pipelines — the metataxonomic workflow and the **(Metatranscriptomics Toward Ecosystem ResponseS) MeToReS** metatranscriptomic pipeline — that convert the resulting reads into biological information. Deliverable D3.3, described here, provides the data and computational infrastructure on which those pipelines are executed and on which the data they produce are stored, managed and made available; the biological results obtained by running these analyses on the MOBILES samples will be reported in deliverable D3.4. In this sense, D3.3 is the platform that operationalises, at scale and reproducibly, the analyses specified in D3.2.

The remainder of this document describes the architecture of this platform: an event-driven, modular and scalable design in which a message broker collects analysis requests and an orchestrator dispatches them to the appropriate bioinformatics pipeline while optimising the use of computing and





MOBILES

storage resources. It then details the hardware infrastructure hosted at the Public University of Navarra (UPNA), the virtualisation and reproducible-environment strategy adopted to guarantee software isolation and long-term maintainability, and the two complementary operational modes — DevOps access for developers and administrators, and a broker-mediated standard access mode for researchers — through which the platform is used. The overarching goal is a robust, scalable and user-friendly environment that makes reproducible metagenomic analysis available to specialist and non-specialist researchers alike, and that remains sustainable beyond the lifetime of the project.

One of the analytical pipelines this infrastructure is designed to support is MeToReS, a Snakemake-based pipeline for soil metatranscriptomics specifically developed by WP3 MOBILES partners.

2. MOBILES project short description

The National Technical University of Athens (NTUA) is working with another 14 partners from academia, research, and industry to develop prototypes of electronic and organism-based biosensors to monitor organic chemicals, antimicrobial-resistant (AMR) bacteria, and pathogens in water, soil, and air.

The MOBILES project is studying and developing biosensors for detecting heavy metals, antibiotics, pesticides, arsenic, microplastic and nanoplastic. It is also developing genetically modified plants and bacteria for detecting heavy metals, antibiotics, and pesticides, and the use of marine diatoms for monitoring bioplastic degradation.

MOBILES aims are to tackle chemicals, including persistent and mobile (PMCs) pollutants and contaminants of emerging concern (CECs), that degrade the environment. Another severe global health risk is associated with increasing AMR in bacteria. Foodborne pathogens, including *Listeria*, *Salmonella*, and *Campylobacter*, pose significant public health risks and are already monitored. However, current bacterial detection methods for environmental control are slow and require specialized laboratories with trained personnel. Similarly, conventional pollutant detection methods, such as chromatography and mass spectrometry, are accurate but time-consuming and require specialized equipment. State-of-the-art detection methods are unsuitable for constant on-site and real-time monitoring. The long time between sampling and detection reduces the efficiency of public health and environmental protection authorities in implementing effective countermeasures. To tackle this problem, several electrochemical biosensors will be developed within the MOBILES project.

Biosensors are devices that combine biological elements with electronic systems to detect specific pollutants. The MOBILES project enhances these sensors with advanced nanomaterials, significantly improving their sensitivity and reliability. All biosensors will have common basic electronics and functional principles (e.g., an organic ligand able to recognize target pollutants), but they will differ in the biological element employed: (i) aptasensors based on aptamers that recognize bacterial cells or spore surfaces, (ii) electronic noses for detecting and quantifying volatile organic compounds (VOCs)





MOBILES

produced by bacteria, (iii) genosensors for detecting genes involved in antibiotic resistance, and (iv) interdigital capacitors functionalized with aptamers for estradiol, a member of CECs family.

Continual threats (such as industrial pollution and the overuse of drugs and pesticides) to sources of drinking water require real-time solutions for wide-ranging water monitoring systems to detect toxicants such as heavy metals, pesticides, and antibiotics. Conventional methods are limited in their ability to detect sub-lethal concentrations of active antibacterial compounds. The damage caused by the activity of an antibacterial agent or pesticide may stimulate different biological mechanisms of bacterial repair. Each antibiotic and/or pesticide triggers specific cellular pathways, mechanisms, and targets within the bacterial cell. This specific biological response, enabling the detection of antibiotics and pesticides using microorganisms, is being investigated in the MOBILES project through the use of genetically modified bacteria to detect toxic pollutants in water. For detecting heavy metals (cadmium, chromium, lead, mercury) in water, MOBILES is developing a flow-through device for continuous monitoring using biological systems (genetically modified bacteria) combined with an optical sensor and flow unit.

Highly toxic arsenic pollution can come from various sources, including industrial activities, mining, and even natural processes. Water and food contaminated by arsenic can cause serious health problems, including cancer and heart disease. For detecting arsenic pollution in soil and groundwater, the MOBILES project is developing genetically modified plants that change colour when arsenic is present in the soil or water used to grow them. The project will conduct safety evaluations to ensure that the genetically modified organisms and developed devices have minimal environmental impact. A specific work package (WP4) is dedicated to evaluate the effects of genetically modified organisms on other organisms and the environment. Safety tests and environmental impact of MOBILES organisms are performed in laboratory using EFSA guidelines.

Microplastic and nanoplastic pollution is raising concerns about its potential impact on human health. The transfer of very small plastics through the trophic chain is a potential source of contamination at all trophic levels. Understanding the distribution, degradation, and life cycle of micro- and nanoplastics in the marine environment is limited by the intrinsic difficulties of current techniques for detecting, quantifying, and chemically identifying small particles in liquids. The MOBILES project is addressing this challenge by utilizing marine diatoms. Diatoms are known for their resilience and adaptability, making them ideal candidates for studying the biodegradation of bioplastics in marine environments. Preliminary studies have shown promising results, indicating that diatoms not only survive in environments containing bioplastics but also contribute to their biodegradation.

In addition to the development of sensors, MOBILES will undertake comprehensive metagenomic analysis, profiling the microbiota of polluted areas across Europe. This work will uncover gene clusters and reveal genetic diversity, enabling a deeper understanding of microbial functions. These insights will provide genetic markers to facilitate rapid evaluation of soil and land health. Two annual sampling rounds are planned for at least two years, and sample collection will be conducted at different locations to target microbiota related to specific pollution types: Greece for urban wastewater contamination, Poland for heavy metal pollution, Cyprus for microplastics and plastics,





MOBILES

France for agriculture and animal farming, Italy for arsenic, and Germany for chemicals and heavy metals from former mining activities. Genomic and transcriptomic data will be analysed, visualized, and interpreted using bioinformatic tools and soil metagenomic web-based platform specifically realized by MOBILES partners. The project's data storage, located in Spain, will be connected to other well-known genomic databases in order to provide a wide range of information.

The biosensors will be rigorously tested with real-world samples from polluted sites to validate their environmental performance.

MOBILES workplan is organized in 6 WPs listed in Table 1. Management actions and collaboration with other EU funded projects are implemented within WP6 while disseminations, exploitation and communication (DEC) activities are grouped within WP5. WP1-WP2 deal with interconnected scientific and technological activities to develop electrochemical biosensors to detect specific pollutants, and organism-based biosensor to monitor other typology of selected pollutants. In WP3 an extensive metagenomic analysis will be performed in order to enable searches for diverse functionalities across multiple gene clusters in polluted and not polluted areas across Europe. In WP4 the safety of all genetically modified bacteria and plants will be tested using standard EU procedures (e.g., EFSA guidelines for genetically modified organisms), and a pre-industrial design of the various biosensors will be provided along with stability (shelf-life) tests.

Table 1. MOBILES WPs list

WP	Work Package Title	Lead Name	Start Month	End month
1	Electronic biosensors for environmental monitoring	INRAE	1	36
2	Detection of pollutants via biotic sensors	UR	1	36
3	Metagenomics database and fully-sequenced polluted soil microbiota	CNR-ISAFO	1	42
4	Environmental performance and safety of developed organisms, and packaging of sensor devices	RICPA	10	42
5	Dissemination, exploitation and communication of project outcomes	GG	1	42
6	Project Management and Coordination	NTUA	1	42





MOBILES

3. Introduction: Genomic Databases

Genomic databases are a key component of modern bioinformatics infrastructure. They provide centralized repositories for the storage, organization, annotation, and dissemination of genomic information. They contain a wide variety of biological data, including genome sequences, gene annotations, protein sequences, genetic variants, transcriptomic profiles, metagenomic assemblies, and associated metadata.

The rapid advances in next-generation sequencing (NGS) technologies have led to an unprecedented growth in genomic data production. Large-scale sequencing initiatives have generated petabytes of genomic information. Consequently, genomic databases have evolved from simple sequence repositories into complex platforms that integrate heterogeneous biological datasets and sophisticated analytical tools.

From a technical perspective, genomic databases can be categorized into primary, secondary and specialized databases. Primary databases store raw experimental data, including nucleotide sequences and sequencing reads (GenBank, the European Nucleotide Archive, and the DNA Data Bank of Japan). Secondary databases provide value-added information through computational processing and annotation of primary datasets (UniProt and Ensembl). Specialized databases focus on particular taxa, diseases, functions, or research areas, including databases dedicated to microbial genomes, cancer genomics, or antibiotic resistance genes. In metagenomic studies, databases such as GTDB, SILVA, RefSeq, and Kraken2 libraries play critical roles in taxonomic classification and functional annotation.

While a defining feature of genomic databases is the integration of biological sequence data with metadata describing sample origin, experimental protocols, environmental conditions, taxonomic classifications, phenotypic characteristics, and clinical information, the construction and maintenance of these complex and dynamic genomic databases present numerous technical and organizational challenges. The quality and completeness of metadata directly influence the scientific value of a database because downstream analyses often depend on contextual information to interpret genomic findings.

Several challenges must be addressed, including data volume and quality, data integration and interoperability, annotation accuracy, metadata standardization, taxonomic and functional classification, and computational and financial costs.

One of the most significant challenges is the enormous volume of genomic data generated by modern sequencing technologies and, by extension, the scalability of those databases. Storing, indexing, and querying these datasets requires highly scalable computational infrastructures capable of handling very high volumes of data (terabytes or petabytes), distributed storage systems, cloud computing platforms, and high-performance computing resources. While researchers require rapid access to specific sequences, taxonomic groups, genomic regions, functional annotations and other features, efficient data retrieval becomes increasingly difficult as databases expand and sophisticated indexing strategies and optimized search algorithms are required.





MOBILES

Genome assemblies and annotations may contain sequencing errors, assembly artifacts, contamination, incomplete regions, or incorrect gene predictions. Low-quality data can propagate through downstream analyses and affect the reliability of scientific conclusions. Functional annotation often relies on sequence similarity searches against existing databases, and errors in reference databases may lead to incorrect gene assignments. In metagenomics, where many organisms are poorly characterized or entirely novel, annotation uncertainty can be particularly high. The continuous improvement of annotation pipelines and expert curation processes is therefore essential to maintain database reliability.

The usefulness of genomic databases depends on sequence quality and on the consistency and completeness of metadata. Data submissions frequently use different formats, terminologies, units, and classification systems. The lack of standardized metadata complicates large-scale comparative studies and automated analyses. Ensuring compliance with metadata standards remains a persistent challenge for database developers.

Genomic databases must continuously adapt to evolving biological knowledge. Taxonomic classifications are regularly revised as new genomes become available and phylogenetic analyses improve. Similarly, functional classifications evolve as novel genes, pathways, and protein families are discovered. Maintaining up-to-date annotations while preserving compatibility with previous database versions requires substantial computational and curatorial effort. Frequent database updates can also affect the reproducibility of analyses when different studies rely on different versions of reference datasets.

Modern biological research relies on the integration of multiple omics datasets, including genomics, transcriptomics, proteomics, metabolomics, and metagenomics. Combining heterogeneous data types requires interoperable database architectures, common ontologies, and standardized interfaces. However, different databases frequently employ distinct data models and annotation systems, creating barriers to interoperability. Ensuring that information can be exchanged seamlessly between repositories remains a major challenge in bioinformatics.

Finally yet importantly, building and maintaining genomic databases requires significant computational resources and long-term financial investment. Database operators must support data storage, backup systems, indexing services, web interfaces, application-programming interfaces (APIs), annotation pipelines, distributed computational and storing resources and operational support. As database size increases, infrastructure costs grow substantially. Long-term sustainability remains a concern for many specialized databases developed through short-term research projects. The data infrastructure for metagenomics analysis here described is an event-driven and scalable platform that decouples request submission, workflow orchestration, computation, and reporting for metagenomics analysis. The architecture relies on a message broker (queue system) to collect incoming analysis requests and an orchestrator to assign each request to the most appropriate bioinformatics pipeline while optimizing computing resources (memory, CPU, and GPU) and storage.





MOBILES

4. MOBILES platform overview

The data infrastructure for metagenomics analysis here described is an event-driven and scalable platform that decouples request submission, workflow orchestration, computation, and reporting for metagenomics analysis. The architecture relies on a message broker (queue system) to collect incoming analysis requests and an orchestrator to assign each request to the most appropriate bioinformatics pipeline while optimizing computing resources (memory, CPU, and GPU) and storage.

This document describes the hardware and software architecture that provide a robust platform for high-throughput metagenomics analysis where a broker reliably collects requests and an orchestrator dynamically dispatches them to specialized computational pipelines that generate biological insights and automated reports. One example of the analytical workflows the platform is designed to support is MeToReS, a Snakemake-based pipeline for soil metatranscriptomics that integrates individual sample assembly, taxonomic classification, differential expression analysis, antimicrobial resistance gene annotation, and environmental chemistry integration. **This pipeline has been specifically designed and implemented for the MOBILES project.** This document also describes the operational workflow, the computation mechanism, and the data storage strategy.

The result is an extensible and modular platform that receives requests for soil sample analysis which organizes, schedules and executes them intelligently according to the availability of computing and storage resources. The analysis is scalable not only in terms of IT resources but also allows for the introduction of new pipelines, new modules/tasks/components into existing pipelines, the comparison of samples against different reference databases, or the presentation of results in various ways, as needed.

The goal is for researchers to be able to specify, using a simple configuration file, the analysis they wish to obtain from the corresponding soil sample, and for the platform to perform that analysis in a way that is transparent to the researcher. The platform notifies researchers via events regarding the progress and availability of the requested analyses.

4.1. MeToReS Pipeline

MeToReS is a Snakemake-based pipeline developed for soil metatranscriptomics, illustrating the type of analytical workflow the platform is designed to support. Starting from raw paired-end sequencing reads, MeToReS performs quality control, adapter trimming, and ribosomal RNA removal, followed by individual per-sample assembly and the construction of a pooled, non-redundant gene catalogue. Genes are quantified, taxonomically classified, and tested for differential expression across experimental groups. Optional modules extend the core analysis with functional annotation, antimicrobial resistance gene detection, and the integration of environmental chemistry measurements to identify associations between gene expression, resistome composition, and soil contaminant levels. Figure 1 presents the complete MeToReS pipeline architecture, illustrating the input, processing tool, and output associated with each analytical stage.





Figure 1. MeToReS pipeline architecture: analytical stages, tools, and outputs.

5. Hardware infrastructure

The metagenomics analysis platform is deployed on a high-performance rack-mounted server (see Figure 2) specifically designed to support computationally intensive bioinformatics workloads, including large-scale sequence analysis, machine learning applications, and GPU-accelerated processing. The system's technical specifications are summarized in Table 2. The hardware is hosted in the data center of the Public University of Navarra (UPNA), which provides a highly reliable operational environment through redundant power supplies, uninterruptible power systems, regular data backup procedures, and redundant Internet connectivity. The resulting infrastructure combines computing, storage, networking, and GPU acceleration resources in a balanced architecture that delivers high performance, reliability, and scalability to support current and future analytical requirements.

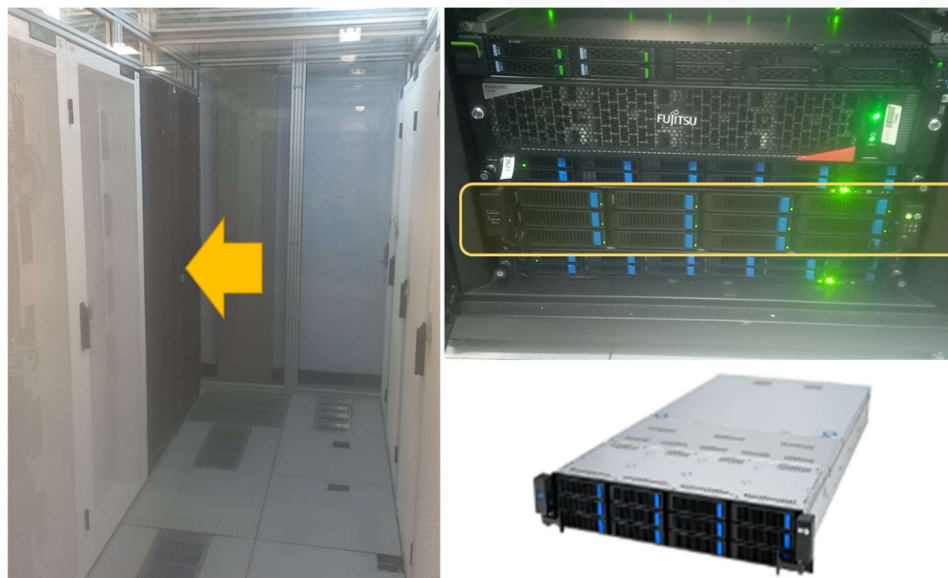


Figure 2. Hardware device (right) and location (left) in the data center of the Public University of Navarra.

Table 2 Hardware infrastructure components.

Computing node	
1x	Dual Server Platform AMD EPYC™ 9004 Series 1+1 RPSU 8x NVMe GPU
2x	AMD EPYC 9454 2,75Ghz 48 cores processor
512 GB	DDR5 4800Mhz ECC RAM Memory
Storage	
2x	512GB SSD SATA3 6Gb/s [Operating system storage]
4x	3.84TB SSD Pcie NVMe U2 [Computing storage]
8x	8TB HD SAS 12Gb/s 7.200rpm [Data storage]
1x	Broadcom MegaRAID 9560
Network	
2x	1Gb Ethernet
2x	10Gb SFP+
GPU	
4x	NVIDIA RTX 6000 ADA 48GB GDDR6 ECC
Description	
Asus RS720A-E12-RS12 (90SF02E1-M00BZ0). 2 x Socket SP5 (LGA 6096) AMO EPYQM 9004 Series. MaxTDP: 300W. Memory Total Slots: 24 (12-channel per CPU). Maximum up to 3072GB per socket. Memory Type: DDR5 4800/4400 RDIMM/ 3DS RD IMM. Expansion Slots Up to 9 PCIe Gen5 slots + 1 x internal RAID slot 6 x PCIe Gen5 x8 or 3 x PCIe Gen5 x16 (FHFL) 2 x PCIe Gen5 x8 or 1 x PCIe Gen5 x16 (FHFL) 1 x PCIe Gen5 x16 (LPHL) 1 x PCIe Gen4 x8 (LP, internal). Drive Bays: 12 x3.5" Hot-Swap drive bays 8NVMe/SATA/SAS* + 4SA TA/SAS*, *SAS support only fi-om optional SAS HBA/RAID card, 2 x M.2 connectors (SATA 6Gb/s & PCI-E Gen4 x4 link). Networking: 1 x Dual Port Intel X710-A T2 IOG LAN Controller + 1 x Mgmt LAN. Graphic As peed AST2600 64MB. OnBoard ASMBI 1-iKVM for KYM-over-IP. Dimensions 840mmx449mmx88.1mm. (2U) Power supply 1+1 Redundant 2600W 80 PLUS Titanium. NVIDIA Ada Lovelace Architecture Process Size: 4nmNVIDIA Custom Process I TSMC Die Size: 608.4 mm Transistors: 76.3 Billion CUDA Cores: 18176. Tensor Cores: 568 Gen-4 RT Cores: 142 Gen-3 GPU Memory: 48 GB GDDR6 with ECC Memory Interface: 384-bit Graphics Bus PCI Express Gen 4.0 x 16 Display. Connectors DP 1.4 (4) Form Factor: 4.4" H x 10.5" L, Dual-Slot Thermal Solution: Active Fans ink	



MOBILES

AR/MR/VR/XR Capable Yes NVIDIA 3D Vision and 3D Vision Pro Support via Optional 3-pin mini-DIN Frame Lock: Compatible with NVIDIA Quadro Sync II Maximum Power Consumption 300 W NVENC. NVDEC 3x I 3x + A V1 Encode and Decode

This hosting environment ensures low latency communication, high bandwidth availability, reliable network performance, and efficient data transfer capabilities. Although low network latency is not a critical requirement for this application—given the large size of the datasets and analysis results, as well as the substantial computational time required for processing—the other network performance characteristics are essential for maintaining service reliability and operational continuity.

To support uninterrupted operation, the infrastructure incorporates automation and monitoring mechanisms, complemented by the expertise of qualified UPNA personnel responsible for system maintenance and service continuity. The data center further provides controlled environmental conditions, including dedicated cooling systems and appropriate humidity regulation, to ensure optimal hardware performance and longevity.

Virtualization technologies have been adopted to facilitate both the rapid deployment of services and the flexible reconfiguration of the infrastructure.

Scalability was a key consideration during the hardware selection process. The chosen infrastructure allows for future expansion of computational and storage resources—including memory, CPU, GPU, and storage capacity—as well as upgrades to network capabilities, ensuring that the platform can accommodate increasing workloads and evolving service requirements.

The platform is based on an ASUS RS720A-E12-RS12 2U enterprise server equipped with dual AMD EPYC™ 9004 Series processors. The system incorporates two AMD EPYC 9454 CPUs, each providing 48 cores operating at 2.75 GHz, for a total of 96 physical cores and 192 processing threads. This multi-core architecture enables the parallel execution of computational pipelines, supporting multiple analyses simultaneously while reducing overall processing times for large metagenomic datasets.

To support memory-intensive bioinformatics applications, the server is equipped with 512 GB of DDR5 ECC RAM operating at 4800 MHz. Error-Correcting Code (ECC) memory enhances system reliability by detecting and correcting memory errors, which is particularly important in long-running scientific computations where data integrity is critical.

Storage is organized into three logical tiers that separate operating system, computation, and data storage workloads. Two 512 GB SATA SSDs provide dedicated storage for the operating system (Linux Ubuntu 24.04) and virtualization environment (Proxmox 8.3.4), enabling fast system boot and administrative operations. Four 3.84 TB PCIe NVMe U.2 SSDs are devoted to high-performance computing storage supplying high-speed storage for temporary datasets, processing pipelines, intermediate results, and software containers. The NVMe interface provides significantly higher throughput and lower latency than traditional storage technologies, accelerating data-intensive analysis tasks. Finally, long-term data storage is provided by eight 8 TB SAS hard drives for storing metagenomic samples, reference databases, and analysis results. These enterprise-grade disks offer a cost-effective balance between storage capacity and reliability. Storage management is





MOBILES

handled by a Broadcom MegaRAID 9560 controller, which provides hardware RAID 6 functionality to improve data protection, storage reliability, and fault tolerance.

The infrastructure also includes a GPU acceleration layer, available to support GPU-enabled workloads as analytical requirements evolve. The server integrates four NVIDIA RTX 6000 Ada Generation GPUs, each equipped with 48 GB GDDR6 ECC memory, 18,176 CUDA cores, 568 fourth-generation Tensor Cores and 142 third-generation RT Cores. Together, the four GPUs provide 192 GB of aggregate GPU memory and substantial parallel computing capability. This acceleration layer is available to support GPU-accelerated sequence analysis and genomic classification workflows. The availability of multiple GPUs also enables concurrent execution of independent analyses or distributed workloads.

High-performance data connectivity and communication management is provided by means of multiple network interfaces, which include two 1 Gigabit Ethernet ports for system administration, monitoring, and standard service operations, and two 10 Gigabit SFP+ interfaces dedicated to high-throughput data transfers between the platform, institutional storage systems, and external users. This architecture ensures enough bandwidth for transferring large sequencing datasets and analysis results while maintaining network segmentation between management and production traffic.

Since the infrastructure is designed to operate continuously in a production environment, service reliability is enhanced through dual 1+1 redundant 2600 W Titanium-rated power supplies, hot-swappable storage bays, ECC memory, enterprise-grade RAID storage management and remote management capabilities through the integrated ASMB11-iKVM controller. These features minimize downtime and facilitate maintenance operations without interrupting service availability.

The platform is designed to run in a virtualized environment, so it requires certain hardware features, such as a large number of CPU cores, high memory capacity, and PCIe Gen5 expansion capabilities. Virtualization enables the deployment of independent services and analysis environments while simplifying administration and resource allocation. The platform supports up to 3 TB of RAM per processor socket, additional PCIe Gen5 expansion cards, and future storage upgrades. Consequently, the platform can accommodate increases in computational demand, storage requirements, and AI-accelerated workloads without requiring a complete infrastructure replacement.

Overall, the proposed architecture provides a robust, highly parallel, and scalable computing environment capable of supporting current and future metagenomics analysis requirements while ensuring high availability, reliability, and efficient handling of large biological datasets.

6. Platform virtualization

To improve infrastructure flexibility, availability, and maintainability, the platform has been deployed on a virtualized environment consisting of a Proxmox 8.3.4 hypervisor and multiple virtual machines (Linux Ubuntu 24.04). This architecture abstracts services from the underlying hardware, allowing computational resources such as CPU, memory, storage, and networking to be allocated dynamically according to workload requirements. In turn, the various services in the pipelines designed and developed to perform metagenomic analyses are based on dedicated Conda environments.





MOBILES

Virtualization provides significant advantages in terms of data protection and backup management since virtual machines can be backed up at the image level (including the operating system, applications, configurations, pipelines and associated data). This approach simplifies the implementation of automated backup and/or snapshot procedures, ensures that complete service environments can be restored rapidly in the event of data loss or system corruption, and reduces downtime.

The virtualized infrastructure also contributes to fault tolerance and service continuity. By isolating services into independent virtual machines and dedicated software environments, failures affecting one component do not necessarily affect the rest of the platform. Furthermore, virtualization platforms facilitate the migration, replication, and restoration of virtual machines, reducing service interruptions and improving overall system resilience.

From an operational perspective, the virtualization architecture followed simplifies system administration and maintenance. This is very important because the goal is to ensure the platform's sustainability and resilience beyond the duration of the project. Hardware resources can be consolidated on a single physical server while maintaining logical separation between services. This reduces infrastructure complexity, improves resource utilization, and allows maintenance activities to be performed more efficiently. Routine administrative tasks, such as provisioning new services, modifying resource allocations, or migrating workloads, can be carried out without requiring physical intervention.

Finally, virtualization facilitates the deployment of updates and service evolution. Software upgrades, configuration changes, and the introduction of new applications can be tested in isolated environments before being applied to production systems. The use of snapshots enables administrators to quickly revert to a previous state if an update introduces unexpected issues, thereby reducing operational risks and ensuring greater service reliability.

Metagenomic pipelines typically integrate multiple bioinformatics tools, databases, scripting languages, and software libraries into a single computational workflow. Common tasks such as quality control, taxonomic classification, assembly, gene prediction, and functional annotation often rely on software developed by different groups and distributed with distinct dependency requirements. Dependency management is a critical challenge. Different applications frequently require different versions of the same library or runtime. Installing these dependencies globally can lead to version conflicts, compatibility issues, and unstable execution environments. As a result, software installation, updates and environment management become significant challenges in metagenomic analyses. Docker¹ and Conda² are two widely adopted solutions that simplify the implementation, deployment, and maintenance of such pipelines.

¹ D. Merkel, "Docker: lightweight linux containers for consistent development and deployment", Linux Journal, vol: 2014, num: 239, pp: 2, 2014.

² CONDA. Computer software. July, 2026. Web. <https://github.com/conda/conda>





MOBILES

6.1. Docker vs Conda

In developing this platform, two possible approaches were considered: the use of Docker containers and the use of the Conda development environment. On this platform, we have chosen to use Conda environments because they offer greater simplicity in terms of management, integration, and orchestration.

Docker addresses the above limitations (software installation, setup, updates...) through containerization. A Docker container packages not only the bioinformatics applications and their dependencies but also the operating system libraries and runtime environment required for execution. Consequently, a containerized metagenomic pipeline can produce highly consistent behavior regardless of the host system on which it is executed. This level of isolation substantially enhances reproducibility, which is a critical requirement for large-scale microbiome studies, regulatory analyses, and research publications. Once a container image has been built and validated, users can execute the entire metagenomic workflow without performing complex software installations. Collaborators need only pull the container image and run the workflow, greatly reducing configuration effort and minimizing the risk of dependency conflicts. Container registries facilitate software distribution by providing centralized repositories for version-controlled images. Docker is also particularly advantageous in cloud computing environments, where computational resources may be provisioned dynamically. Containers can be deployed consistently across local workstations, cloud platforms, and high-performance computing (HPC) systems, enabling scalable analyses without requiring substantial modifications to the workflow. However, Docker container images may consume several gigabytes of storage because they package complete software stacks and operating system components. Building and maintaining container images also requires additional expertise in container technologies. Furthermore, some HPC systems restrict direct Docker execution due to security concerns.

Conda addresses these limitations by creating isolated environments where each project maintains its own dependency stack. A Conda environment is an isolated software environment managed by the Conda package and dependency management system. Conda allows users to create project-specific environments without affecting the host operating system. It encapsulates a specific interpreter version, installed packages, libraries, and their dependencies, enabling applications and development workflows to operate independently of one another on the same system. Through repositories such as Bioconda, thousands of bioinformatics packages can be installed with minimal effort, eliminating the need for manual compilation and reducing installation errors. This capability is particularly valuable during pipeline development, when tools are frequently added, removed, or upgraded. Researchers can rapidly create multiple environments to test alternative software versions or workflow configurations while maintaining separation between projects.

Another advantage of Conda is its seamless integration with workflow management systems such as Snakemake and Nextflow. These frameworks can automatically generate and activate Conda environments for individual pipeline processes, ensuring that each analysis step executes with the required dependencies. Such functionality improves reproducibility while simplifying workflow maintenance. Furthermore, Conda environments are relatively lightweight compared with full containers, making them attractive in shared computing environments where storage resources may be limited.





MOBILES

Despite its convenience, Conda environments do not completely isolate software from the underlying operating system. Differences in system libraries, operating system versions, or hardware architectures can occasionally lead to inconsistencies when a pipeline is transferred between computational infrastructures. Although Conda significantly improves dependency management, reproducing an identical execution environment across multiple institutions or computational platforms may still be challenging.

6.2. Conda's election

For all of the above reasons, and given that there is no intention to transfer pipeline computations between different computational infrastructures, this platform has opted to use Conda instead of Docker technology. Although a combined strategy that exploits Conda for development and Docker for deployment often represents an effective and reproducible solution, this platform has opted to use Conda exclusively to make it easier for users who are not experts in virtualization and cloud resource orchestration to use and manage the infrastructure.

Rather than containerizing each pipeline component with Docker, as some well-known pipelines do, software isolation in this platform is achieved through dedicated Conda environments, each of which encapsulates the specific interpreter, libraries, and dependency versions required by a given set of tools. This distinction matters in practice: certain bioinformatics tools impose unusually narrow version constraints on their own dependencies, which can conflict with the broader, more modern dependency tree required by the rest of the pipeline. For example, in the MeToReS pipeline, the Resistance Gene Identifier (RGI), used for antimicrobial resistance gene annotation, requires an older BLAST and Python toolchain that is incompatible with the DIAMOND-based dependency stack shared by the remaining analysis tools.

Rather than forcing a single, monolithic environment to satisfy every tool's constraints simultaneously, RGI is isolated in its own dedicated Conda environment, resolved dynamically by name at runtime, while every other pipeline component shares a single, consolidated environment. This approach preserves the core benefit generally attributed to containerization, namely fault isolation and freedom from version conflicts, without the additional overhead and maintenance burden of building and maintaining container images for every pipeline component.

6.3. Platform description

Modern metagenomics pipelines typically consist of multiple interdependent analysis steps, including quality control, sequence preprocessing, taxonomic classification, functional annotation, statistical analysis, and report generation. Managing these complex workflows manually is error-prone and difficult to reproduce, particularly when processing large datasets or multiple samples simultaneously. This platform uses Snakemake as its workflow management system to ensure reproducibility, scalability, maintainability, and efficient resource utilization.

Snakemake represents the analysis pipeline as a set of rules connected through explicit input-output dependencies. This approach enables the automatic determination of execution order, ensuring that each analysis step is performed only when its required inputs are available. Furthermore, Snakemake





MOBILES

tracks generated outputs and executes only those tasks that are necessary, avoiding redundant computations and reducing overall execution time. Since researchers may need to perform different analyses on the same soil samples, Snakemake provides version-controlled alongside the analysis software and configuration files. This enables researchers to reproduce analyses consistently over extended periods.

The workflow engine supports parallel execution of independent tasks and can automatically distribute workloads across multiple CPU cores, computing nodes, or accelerator resources. This capability is especially valuable in the proposed infrastructure, which includes high-core-count AMD EPYC processors capable of executing multiple independent analysis tasks concurrently. By scheduling tasks according to resource requirements, Snakemake maximizes hardware utilization and reduces analysis turnaround times.

Snakemake also facilitates integration with environment and dependency management systems, such as Conda, eliminating dependency conflicts and simplifying software deployment. As a result, workflow portability is improved and the effort required for system maintenance is reduced.

From an operational perspective, Snakemake simplifies workflow monitoring, error handling, and recovery. If an execution is interrupted due to hardware failures, software errors, or administrative actions, the workflow can be resumed from the point of failure without repeating successfully completed steps. This checkpoint-based execution model improves reliability and reduces computational waste, particularly when processing large metagenomic datasets that require substantial execution times.



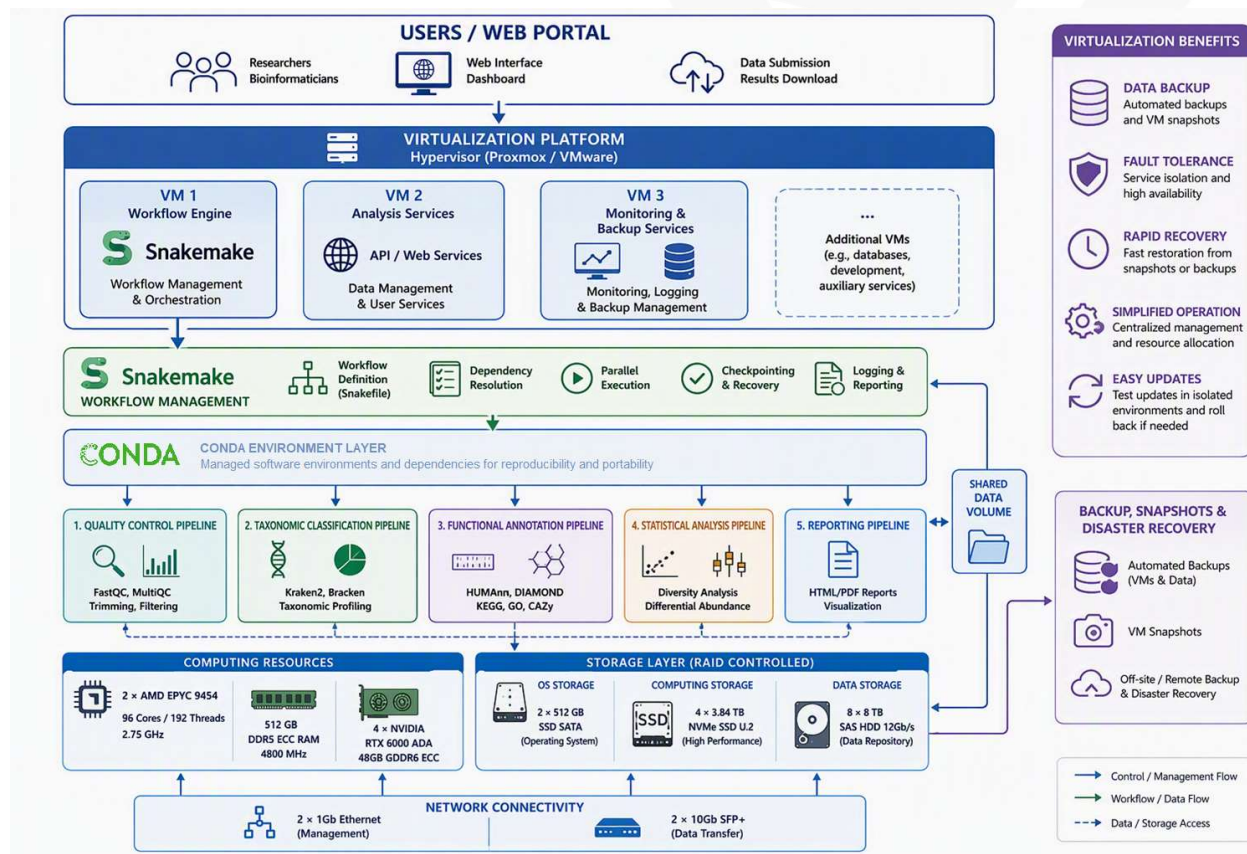


Figure 3. Architecture of the metagenomics analysis platform.

Figure 3 depicts the architecture of the metagenomics analysis platform. The infrastructure is deployed on a virtualized environment managed by a hypervisor. Snakemake orchestrates the execution of bioinformatics workflows, while individual pipeline components are executed within dedicated Conda environments to ensure reproducibility and software isolation. Computational workloads are distributed across multi-core AMD EPYC processors.

Data are stored on a multi-tier storage subsystem combining SSD, NVMe, and SAS technologies, while virtualization enables snapshots, backup services, fault tolerance, and rapid recovery procedures.

Fault recovery is managed at the level of individual Snakemake rules rather than individual containers. Since Snakemake tracks the input-output dependency graph and the completion status of every rule, execution can be interrupted, corrected, and resumed without repeating already-completed analysis steps. For example, after quality control, adapter trimming, and host filtering have completed successfully, an assembly failure due to a misconfigured parameter does not require restarting the analysis from the beginning. Instead, the failed rule is corrected in the workflow configuration, and execution is resumed from the last successfully completed step, with Snakemake automatically determining that upstream outputs remain valid and do not need to be regenerated. From there, the pipeline continues with taxonomic classification, functional annotation, and, finally,



MOBILES

report generation. This checkpoint-based recovery behaviour has been directly validated during the development of the MeToReS pipeline.

7. Operational modes

This section describes the target operational architecture to be implemented by the completion of the project. The platform is designed to support two complementary operational modes designed to accommodate users with different technical expertise and requirements: DevOps access and standard user access.

The coexistence of these two operational modes provides a balance between flexibility and usability. Advanced users can directly manage and evolve the bioinformatics workflows through the DevOps interface, while standard users benefit from a fully automated environment where an orchestrator transparently coordinates Snakemake workflows, Conda environments, virtualized services, storage systems, and computational resources. This dual-access architecture facilitates both the continuous development of the platform and its adoption by non-specialized researchers.

7.1. DevOps Access Mode

The DevOps Access mode is intended for developers and system administrators who require direct interaction with the computational infrastructure. In this mode, users have full access to the different components of the platform, including the Snakemake workflow engine, configuration files, software containers, and individual pipeline stages.

This access level enables users to modify and extend Snakemake workflows —such as MeToReS— develop and test new pipelines, deploy new analysis services, update existing components, debug workflows, and monitor workflow execution from the developer perspective. From the administrator perspective, it provides the ability to configure computational resources and execution parameters, manage the pipeline's Conda-based software environments, and administer the execution environment.

By exposing all workflow components, the DevOps Access mode provides maximum flexibility and facilitates continuous integration, software maintenance, and the incorporation of novel analysis methodologies.

DevOps Access mode is provided through a secure Remote Desktop Protocol (RDP) connection protected by two-factor authentication (2FA). After successful authentication, users are granted access to the Linux virtual machine according to their assigned roles and privileges. This role-based access control mechanism ensures that each user can only perform operations authorized for their responsibilities, such as workflow development, pipeline deployment, system administration, or infrastructure monitoring. The combination of RDP, 2FA, and role-based permissions provides a secure environment for managing the bioinformatics platform while protecting computational resources and sensitive data from unauthorized access. Figure 4 depicts the DevOps access. It is important to note that this access method ensures that the internal Proxmox network remains isolated for security reasons.



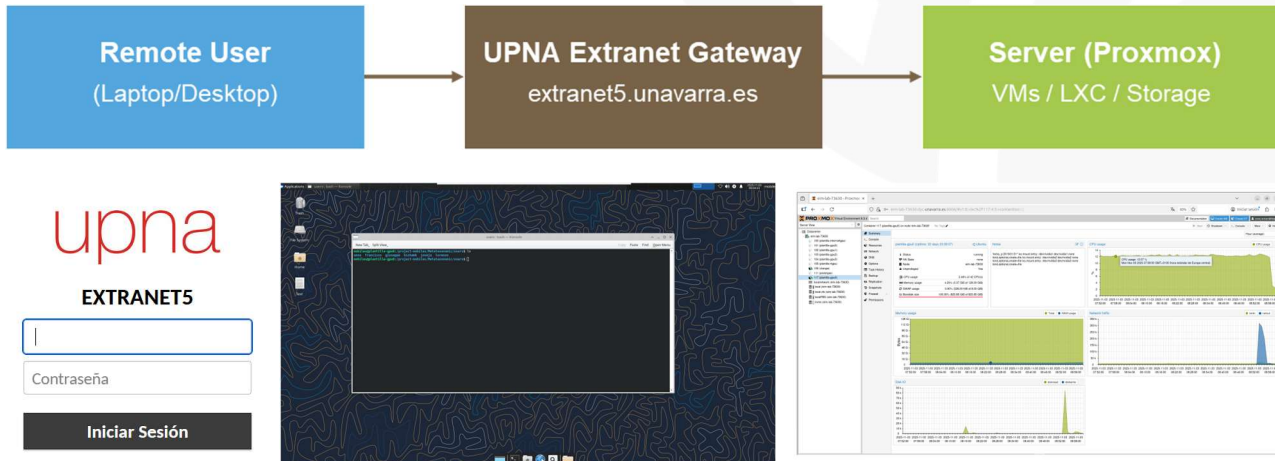


Figure 4. DevOps access. Access flow (top), login validation (left), RDP access to desktop (middle) and Proxmox management console (right).

7.2. Standard User Access Mode

The standard access mode is designed for researchers and end users who require metagenomics analyses without interacting directly with the underlying infrastructure. In this operational model, users access the platform through a web-based interface and interact exclusively with a broker service that provides an abstraction layer between users and the computational infrastructure. The broker receives analysis requests through the web interface or API, validates the submitted data and metadata, and generates the corresponding execution tasks. Once a request has been accepted, the broker determines the appropriate analysis workflow and prepares the execution environment.

The orchestrator then uses the configuration information provided by the broker to orchestrate the entire analytical process. First, it stores the input datasets in the shared storage subsystem and registers the execution request in the platform's internal database. Subsequently, it generates the required Snakemake configuration files and invokes the corresponding workflow. During execution, the orchestrator monitors the status of individual workflow stages, tracks resource utilization, captures execution logs, and manages possible error conditions.

Through its interaction with Snakemake, the orchestrator delegates the execution of each MeToReS pipeline invocation to a dedicated Docker container, within which the pipeline's own Conda-based environments manage individual tool dependencies. Snakemake resolves workflow dependencies and schedules tasks, while the orchestrator supervises the overall execution lifecycle, including job submission, progress monitoring, failure detection, notification generation, and results delivery.

Depending on the computational requirements of each task, the orchestrator allocates CPU cores, memory resources, storage volumes, and, when required, GPU devices. This mechanism prevents resource contention among concurrent users and ensures efficient utilization of the computing infrastructure.



Upon completion of the workflow, the orchestrator collects the generated outputs, stores the results in the data repository, generates execution reports, and notifies the user that the analysis has finished. The user can then access the results through the platform interface without requiring knowledge of the underlying workflow implementation.

By centralizing request management, workflow orchestration, resource allocation, and result delivery, the orchestrator provides a secure and user-friendly interface that hides the complexity of the bioinformatics infrastructure while ensuring reproducible and automated execution of metagenomics analyses. The orchestrator abstracts the complexity of the infrastructure, providing a simplified and user-friendly interface while ensuring that analyses are executed according to predefined policies and resource allocation rules. From the user's perspective, the entire workflow appears as a single automated service, whereas internally the orchestrator coordinates the execution of multiple pipeline stages, containers, virtual machines, storage resources, and computing devices.

Figure 5 shows the different role access to task scheduling, task monitoring/alert notification and visualization/results recovery, while Figure 6 describes the architecture of this access mode.

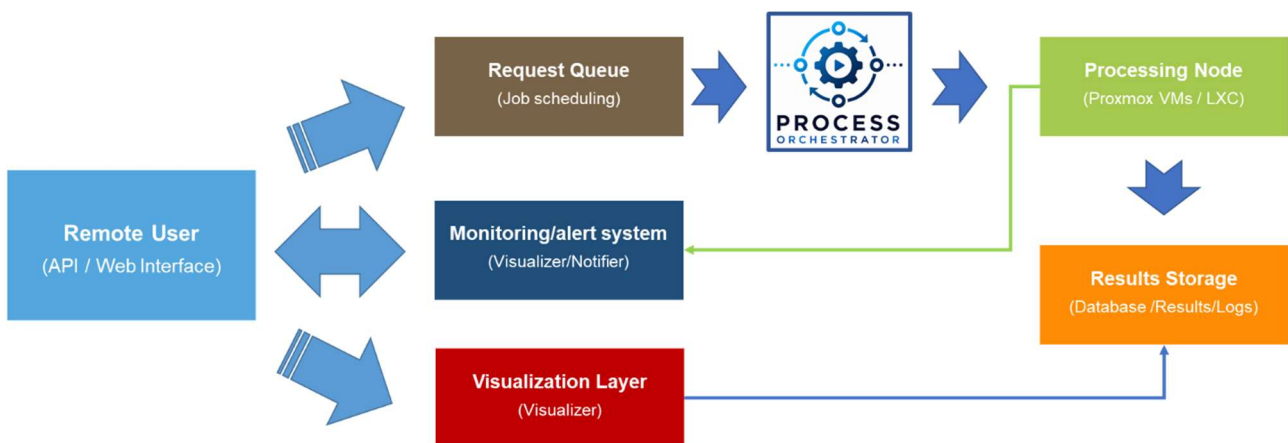


Figure 5: Role-based access.

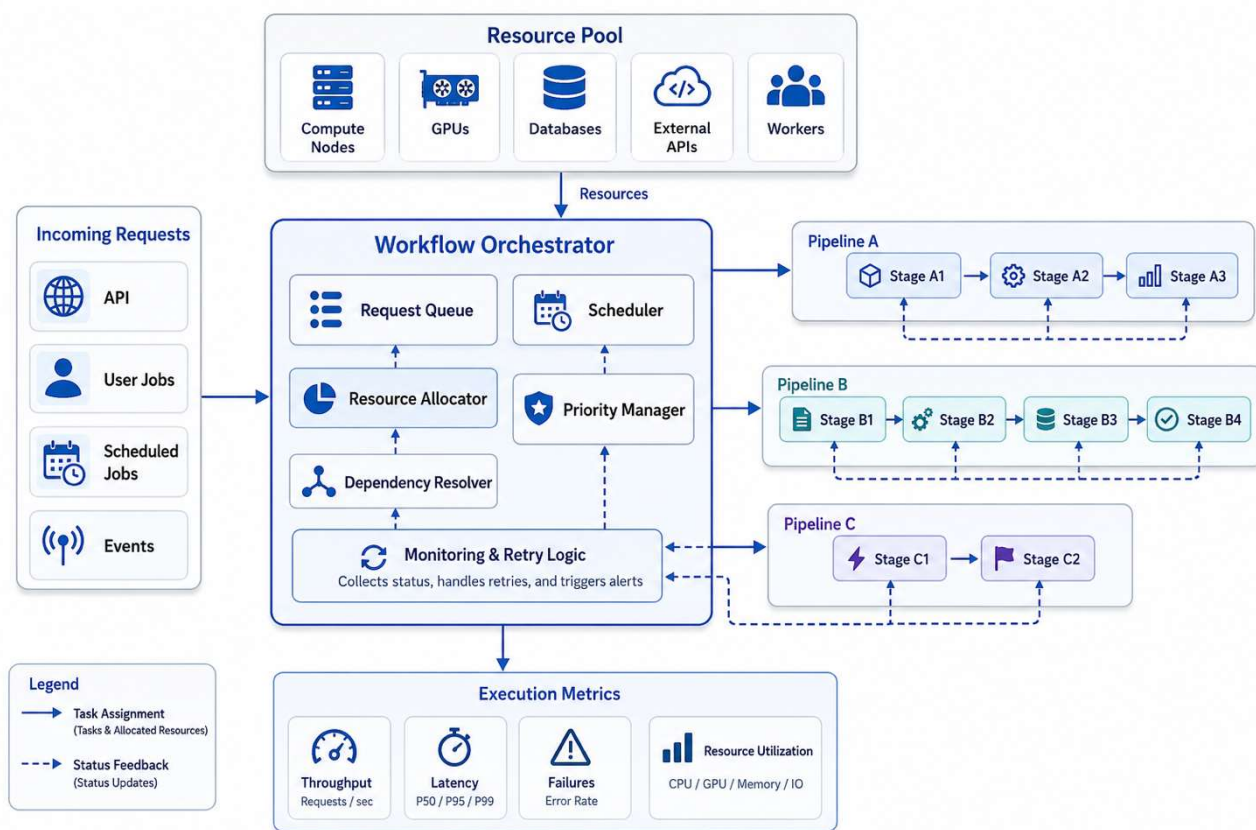


Figure 6: Standard User Access Mode operation diagram.



MOBILES

8. Conclusions

Deliverable D3.3 has presented the design and implementation of the MOBILES data infrastructure for metagenomics analysis. The proposed solution addresses the computational, storage, reproducibility, and management challenges associated with modern high-throughput metagenomic and metatranscriptomic studies by combining a high-performance hardware platform with a scalable and modular software architecture.

The infrastructure deployed provides a robust computational environment based on multi-core processors, high-capacity memory, GPU acceleration capabilities, and a multi-tier storage architecture. These resources enable the efficient execution of computationally demanding bioinformatics workflows while ensuring reliability, fault tolerance, and future scalability. The adoption of virtualization technologies further improves operational flexibility, simplifies system maintenance, and facilitates backup and recovery procedures.

The platform integrates Snakemake workflow management with Conda-based dependency isolation to ensure reproducible, maintainable, and scalable execution of metagenomic analyses. This approach allows complex analytical workflows, such as the MeToReS pipeline developed within the MOBILES project, to be executed efficiently while avoiding dependency conflicts and minimizing administrative overhead. The platform architecture also supports fault recovery and workflow resumption, reducing computational waste and improving operational robustness.

A key characteristic of the proposed infrastructure is its extensibility. Although MeToReS serves as the reference workflow described in this deliverable, the architecture has been designed to accommodate additional metagenomic, metatranscriptomic, and multi-omics pipelines, new reference databases, and emerging analytical methodologies. The modular design ensures that future developments can be incorporated with minimal impact on existing services and workflows.

The platform also introduces a dual operational model that balances flexibility and usability. Advanced users and developers can directly access and modify workflows through the DevOps mode, while researchers without bioinformatics expertise can submit analyses through a simplified interface supported by an orchestration layer that automates workflow execution, resource allocation, monitoring, and result delivery. This approach reduces technical barriers and promotes broader adoption of advanced metagenomic analyses within the MOBILES consortium.





MOBILES

9. List of Figures

Figure 1. MeToReS pipeline architecture: analytical stages, tools, and outputs.

Figure 2. Hardware device (right) and location (left) in the data center of the Public University of Navarra.

Figure 3. Architecture of the metagenomics analysis platform

Figure 4. DevOps access. Access flow (top), login validation (left), RDP access to desktop (middle) and Proxmox management console (right).

Figure 5: Role-based access.

Figure 6: Standard User Access Mode operation diagram.

10. List of Tables

Table 1. MOBILES WPs list.

Table 2. Hardware infrastructure components.

11. List of Abbreviations

Abbreviation	Abbreviation for
2FA	Two-Factor Authentication
API	Application Programming Interface
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DDR	Double Data Rate
DevOps	Development Operations
ECC	Error-Correcting Code
GDDR	Graphics Double Data Rate
GPU	Graphics Processing Unit
IT	Information Technology
MeToReS	Metatranscriptomics Toward Ecosystem ResponseS
NVMe	Non-Volatile Memory Express
PCI	Peripheral Component Interconnect





RAID	Redundant Array of Independent Disks
RAM	Random Access Memory
RDP	Remote Desktop Protocol
RT	Ray Tracing
SAS	Serial Attached Small Computer System Interface
SPF	Sender Policy Framework
WP3	Work Package





MOBILES

12. Project Consortium



www.ntua.gr/en/



www.cnr.it/en



www.inrae.fr/en



www.uniroma1.it/en/pagina-strutturale/home



www.eden-microfluidics.com/ *Terminated consortium membership effective June 2026*



<https://www.unavarra.es/home>



www.en.iung.pl/



www.agri.gov.it/en/home



www.u-bordeaux.fr/en



www.cut.ac.cy/?languageId=1



www.chem.bg.ac.rs/index-en.html



www.mat4nrg.de/



www.tu-clausthal.de/en/



www.grant-garant.cz



www.proakademia.eu/en/

