

Disclaimer:

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



D3.2 – Reference protocols to be used for DNA and RNA-seq

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 D3.2 - Reference protocols for DNA and RNA-seq

Short summary: This deliverable reports the work performed in WP3 Task 3.2: *DNA sequencing and transcriptomic approaches*. In this task the reference protocols, bioinformatic pipelines and computational methods adopted within MOBILES are defined. The document discusses the sequencing-based characterisation of soil microbial communities, covering 16S rRNA amplicon sequencing for metataxonomic profiling and total RNA sequencing (metatranscriptomics) for the analysis of active microbial gene expression.

Due date: 31/08/2026
WP, leader: [WP3, CNR-ISAFOM]
Authors: [WP3, CNR-ISAFOM]

Dissemination Level

PU	Public	☒
SEN	Sensitive	☐

Document history

Version	Date	Name	Chapters edited	Reason for change
V1	01/07/2026	Lorenzo Pin	All	Document drafted
V1.1	09/08/2026	Giuseppe Puglia	All	Document prepared
V1.2	12/08/2026	Angelo Ferraro	All	Document reviewed
V2	20/08/2026	Giuseppe Puglia and Lorenzo Pin	All	Document revised
V2.1	25/08/2026	Angelo Ferraro	All	Document approved



MOBILES

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.....	7
2. MOBILES project short description	8
3. Introduction: Microbiota Soil Analysis	11
4. Scope and Relation to Other Deliverables.....	12
5. Nucleic Acid Sequencing.....	12
5.1. From Quantification to Library Preparation.....	13
5.2. 16S rRNA Amplicon Sequencing (Illumina, PE250)	13
5.3. Long-read 16S Amplicon Sequencing (Oxford Nanopore MinION).....	14
5.4. Metatranscriptomic Sequencing (Illumina NovaSeq, PE150)	15
5.5. Comparison of the Sequencing Strategies	16
6. 16S rRNA Amplicon Sequencing for Metataxonomic Analysis.....	16
6.1. Overview	16
6.2. Preliminary Quality Control.....	19
6.2.1. Global Quality Control (FastQC / MultiQC)	19
6.2.2. Sample-level Preliminary Analysis (seqkit).....	19
6.3. Adapter and Primer Trimming (Cutadapt)	19
6.4. Post-trimming Quality Control	20
6.5. DADA2 Pipeline: Denoising and ASV Generation.....	20
6.5.1. Quality Filtering	20
6.5.2. Error Rate Learning.....	20
6.5.3. Sample Inference (Denoising)	21
6.5.4. Paired-end Merging.....	21
6.5.5. Sequence Table Construction and Chimera Removal	21
6.5.6. Taxonomic Assignment	21
6.6. Pipeline Summary and Outputs	21
6.7. Statistical Analysis of the Metataxonomic Data	22
6.7.1. Data Curation and Filtering	22
6.7.2. Normalisation	22
6.7.3. Alpha Diversity.....	22





MOBILES

6.7.4.	Beta Diversity	23
6.7.5.	Differential Abundance	23
6.7.6.	Pollution-associated Taxa and Community Composition.....	24
7.	MeToReS: an End-to-End Snakemake Pipeline for Metatranscriptomic Analysis of Toxicity and Antimicrobial Resistance	24
7.1.	Overview	24
7.2.	Pipeline Workflow.....	25
7.3.	Quality Control	28
7.3.1.	Initial Quality Control	28
7.3.2.	Adapter Trimming and Quality Filtering.....	28
7.3.3.	Ribosomal RNA Removal	28
7.4.	Assembly and Gene Catalogue Construction.....	28
7.4.1.	Per-sample Assembly and ORF Prediction	29
7.4.2.	Assembly Quality Control	29
7.4.3.	Gene Catalogue Construction and Taxonomic Classification.....	29
7.5.	Gene Expression Quantification.....	29
7.6.	Differential Expression Analysis	30
7.7.	Functional Annotation	30
7.8.	Downstream Analysis.....	30
7.9.	Constrained Ordination (db-RDA).....	31
7.10.	Resistance Gene Annotation (ARG and MRG)	32
7.11.	Chemical Data Integration	32
7.12.	Bioindicator Screening by Partial Correlation.....	33
7.13.	Multivariable Association Modelling (MaAsLin2)	33
8.	Integration of Chemical and Metataxonomic Data for Bioindicator Identification	34
8.1.	Chemical Data and Preparation	34
8.2.	Community-level Relationship with the Chemistry	34
8.3.	Taxon-Element Correlations	35
8.4.	Consensus-anchored, Site-level Screening	35
8.5.	Leverage Diagnostics.....	37





MOBILES

8.6. Cross-season Reproducibility 37

9. Conclusion 38

10. List of Figures..... 39

11. List of Tables 39

12. List of Abbreviations 39





MOBILES

1. Summary

This deliverable describes the reference protocols, bioinformatic pipelines, and computational methods employed for the sequencing-based characterisation of soil microbial communities within the MOBILES project (WP3, Task 3.2). For each nucleic acid it specifies the input requirements, library preparation, sequencing configuration and primary data processing, together with the reference analytical workflows applied to the resulting data. Three complementary sequencing strategies are addressed: (1) short-read 16S rRNA amplicon sequencing on the Illumina platform for metataxonomic profiling of bacterial communities; (2) long-read 16S rRNA amplicon sequencing on the Oxford Nanopore MinION, which recovers the near full-length gene for finer taxonomic resolution; and (3) total RNA sequencing (metatranscriptomics) on Illumina for the assessment of active gene expression in the soil microbiome.

The 16S rRNA amplicon sequencing workflow targets the hypervariable V3-V4 regions of the bacterial 16S gene using paired-end Illumina sequencing. Raw reads are subjected to quality control, primer trimming with Cutadapt, and denoising through the DADA2 framework. The pipeline generates Amplicon Sequence Variants (ASVs) that are taxonomically assigned against the RDP reference database.

The metatranscriptomic workflow is implemented as a Snakemake-based pipeline (MeToReS) that automates all steps from raw FASTQ files to functional annotation and downstream biological interpretation. The pipeline performs quality control and adapter trimming, ribosomal RNA removal, per-sample de novo assembly with a pooled non-redundant gene catalogue, gene prediction, taxonomic classification, expression quantification, differential expression analysis, and functional annotation using a suite of established bioinformatic tools.

Both workflows have been applied to soil samples collected across multiple European sites within the MOBILES sampling campaign (Spring 2025) as described in D3.1. The protocols described herein are intended to serve as reference methods for the MOBILES consortium and to ensure reproducibility of results across sites and time points.





MOBILES

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) 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





MOBILES

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, 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

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





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-ISAFOM	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: Microbiota Soil Analysis

Soil hosts one of the most diverse microbial communities on Earth, the greater part of which cannot be recovered in culture and can therefore be characterised only by sequencing the nucleic acids extracted directly from the matrix. Within the MOBILES project this culture-independent characterisation serves a defined purpose: to describe how the composition and the activity of the soil microbiota change under different types of contamination across European sites, and ultimately to identify microbial taxa and genes that may be proposed as biological indicators of pollution and of antimicrobial resistance. The present deliverable builds directly on D3.1, which described the sampling campaign and the extraction, quantification and quality control of DNA and RNA from the MOBILES soils, and defines how those extracts are converted into sequence data and, through standardised analytical workflows, into the taxonomic and functional information on which D3.4 will draw.

The taxonomic classification of an uncultured community relies on sequencing a phylogenetic marker gene, most commonly the small-subunit ribosomal RNA gene (16S rRNA) in bacteria and archaea. This gene is present in every prokaryote and combines regions that are strongly conserved across all taxa with hypervariable regions whose sequence varies from one taxon to another. The conserved regions provide anchoring sites for universal primers, so that the marker can be amplified from essentially any member of the community, while the hypervariable regions carry the signal that discriminates between taxa. Sequencing the amplified marker — an approach known as amplicon sequencing, or metabarcoding — and comparing the resulting sequences against curated reference databases yields a profile of which organisms are present and in what relative proportions.

Two sequencing technologies, differing fundamentally in the length of the fragment they read, are applied to this marker within WP3. Short-read sequencing on the Illumina platform reads fragments of a few hundred bases with very high per-base accuracy and very high throughput and is the established standard for metabarcoding; because a single read covers only part of the gene, however, the analysis targets a defined window — the V3–V4 hypervariable regions — and the taxonomic resolution it affords generally reaches the genus level. Long-read sequencing on the Oxford Nanopore platform instead reads each molecule along its entire length as it passes through a nanopore, so that a single read can span almost the whole 16S gene; the additional information carried by the full-length marker supports finer assignments, potentially to the species level, at the cost of a lower per-base accuracy. The two approaches are complementary, and both are used here so as to combine the accuracy and depth of short reads with the taxonomic resolution of long reads.

Sequencing DNA reveals which organisms are present, but not which of their functions are actually expressed. This is the domain of metatranscriptomics, the sequencing of the collective messenger RNA of the community, which reports the genes that are actively transcribed under the sampled conditions and therefore reflects the functional response of the microbiota rather than its genetic potential alone. Because the overwhelming majority of the RNA recovered from soil is ribosomal and uninformative for this purpose, it is depleted before sequencing so that the messenger fraction is enriched. The taxonomic profiles obtained from the marker gene and the expression profiles





MOBILES

obtained from the transcriptome are complementary, together describing both the structure of the community and its activity.

Raw sequence reads are not in themselves interpretable: they become taxonomic and functional information only when they are processed through defined bioinformatic workflows and compared against reference databases. The decisions taken at this stage — the trimming and denoising parameters, the reference database chosen for classification, the assembly and quantification strategy, the statistical models applied — shape the results as much as the laboratory protocols do, and they must be fixed in advance if samples processed at different times and by different partners are to remain comparable. For this reason the present deliverable specifies not only the sequencing protocols but also the two reference analytical workflows adopted in WP3: a metataxonomic workflow for the amplicon data and the MeToReS pipeline for the metatranscriptomic data. The data infrastructure that collects, stores and makes accessible the sequence data generated in WP3 — delivered as a document — is the subject of the companion deliverable D3.3.

Taken together, these sequencing technologies and analytical workflows constitute the methodological basis on which the microbial taxa and functions characteristic of polluted European soils will be identified and, in D3.4, proposed as biological indicators of environmental contamination.

4. Scope and Relation to Other Deliverables

This deliverable establishes the procedures, parameters and analytical workflows agreed within WP3, so that samples processed at different times and by different partners were treated consistently and the results obtained across the MOBILES sites remain comparable. The protocols described here were applied to the soil samples collected during the MOBILES Spring 2025 campaign; the complete biological results derived from them will be reported in the forthcoming deliverable D3.4.

The deliverable sits between two others. Deliverable D3.1 precedes it and describes sampling and the extraction, quantification and quality control of the nucleic acids; the acceptance thresholds reported there are the entry criteria for the library preparation described in Section 4 of this document. Deliverable D3.4 follows it and will report the outcomes obtained by applying these protocols: the species composition of the European polluted areas, the differentially expressed genes and putative gene clusters recovered from the metatranscriptomic data, the availability of these data through the project database, and the microbial taxa and genes that can be proposed as biological indicators.

5. Nucleic Acid Sequencing

This section describes the step that connects the quantified nucleic acid extracts to the bioinformatic workflows: their conversion into sequencing libraries and, ultimately, into the raw sequence files that constitute the input of the analyses described in Sections 6 and 7. The three strategies, and the workflow to which each of them feeds, are compared in Table 2.





MOBILES

5.1. From Quantification to Library Preparation

The quantification and purity measurements reported in D3.1 are the acceptance criteria that determined whether an extract could enter library preparation. Each sequencing strategy specifies minimum thresholds for concentration, total amount and volume and, for RNA, also for purity and integrity. Extracts that did not meet these thresholds were re-extracted, concentrated, or excluded from the corresponding batch, because insufficient or degraded input material propagates directly into the quality of the libraries and hence into every downstream analysis.

The requirements differ markedly between the two nucleic acids. DNA is chemically stable and is required in modest amounts, because the target region is subsequently amplified by PCR. RNA is far more labile, is required in larger amounts, and must additionally satisfy purity and integrity criteria, since neither degradation nor co-extracted inhibitors can be compensated for at a later stage.

5.2. 16S rRNA Amplicon Sequencing (Illumina, PE250)

Amplicon sequencing was performed on DNA extracted from the soils sampled within the project. The minimum requirements for the extract were a concentration of at least 1 ng/μL, a total amount of at least 200 ng and a volume of at least 20 μL; as indicative starting material, these amounts are typically obtained from 1–2 g of soil.

The V3–V4 hypervariable regions of the bacterial 16S rRNA gene were amplified with the 341F/806R primer pair (see section 6.3 for sequences), carrying Illumina overhang adapter sequences at their 5' ends so that the sequencing adapters could be added in a subsequent indexing reaction. Amplification used a high-fidelity polymerase, in order to minimise polymerase errors that would otherwise be indistinguishable from genuine biological variants during denoising, and yielded an amplicon of approximately 460 bp. The primer pair used at the bench is the same one removed *in silico* by the trimming step of the metataxonomic workflow, and the expected amplicon length is what determines the read-length filters and the overlap requirement applied during denoising.

Library construction proceeded through bead purification of the amplification products (AMPure XP), a second limited PCR that attached the dual indices (i5 and i7) and the P5/P7 sequencing adapters (standard Illumina dual-index adapters) so that each sample carried a unique barcode combination, and a further clean-up that removed primer dimers. The library was checked for size, expected around 550 bp once adapters are included, and quantified by Qubit or qPCR. Libraries were then normalised to equimolar concentration and pooled, which ensured that the sequencing depth was distributed evenly across the multiplexed samples.

Sequencing was performed on an Illumina instrument, MiSeq or, for larger batches, a NovaSeq, in paired-end mode with 250 bp per read (PE250). Libraries were clonally amplified on the flow cell by bridge amplification and read by sequencing-by-synthesis chemistry with reversible terminators; for dual-indexed libraries the instrument performs four successive reads, in the order Read 1, Index 1,





MOBILES

Index 2 and Read 2. The requested depth is specified per sample and is commonly set at 50,000, 100,000 or 300,000 read pairs. These options are not analytically equivalent: 50,000 read pairs support only a minimal ecological profiling, 100,000 represent the standard choice for community comparison, and 300,000 are appropriate when high resolution and the detection of rare taxa are required.

The raw instrument output consists of BCL files, which were converted into sequence files and assigned to the individual samples on the basis of their index combinations using bcl2fastq or an equivalent tool such as DRAGEN. The output of the sequencing service was therefore a pair of demultiplexed FASTQ files (R1 and R2) per sample, with a typical quality expectation of Q30 above 80–85% and minimal adapter contamination. Because the PE250 configuration is longer than half of the amplicon, the forward and reverse reads overlap, which is the property exploited by the paired-end merging step of the metataxonomic workflow.

5.3. Long-read 16S Amplicon Sequencing (Oxford Nanopore MinION)

In addition to the two short-read strategies described above, a portion of the DNA extracted from the MOBILES samples was sequenced at CNR-ISAFOM on the Oxford Nanopore MinION, a third-generation platform that reads single DNA molecules as they are driven through a biological pore, so that the length of the read is limited by the length of the molecule presented to the pore rather than by the chemistry of the instrument.

Libraries were prepared with the Oxford Nanopore Microbial Amplicon Barcoding kit, which combines the amplification of the 16S rRNA gene with the attachment of sample-specific barcodes, allowing several samples to be pooled on a single flow cell. The essential difference with respect to the Illumina approach lies in the extent of the target: whereas the short-read protocol interrogates the V3–V4 regions alone, the long-read protocol recovers the gene along nearly its entire length, so that a single read spans all hypervariable regions. This carries considerably more discriminating information and can support finer taxonomic assignments, potentially at species level.

The MinION produces raw electrical signals that must be converted into sequence by a computationally intensive basecalling model; for this reason the platform was operated at CNR-ISAFOM, connected to high-performance computing infrastructure equipped with a high-performance graphics card, which performed the basecalling and, on the basis of the sample barcodes, the demultiplexing that yielded the per-sample FASTQ files of long reads. Data production is the scope of the present deliverable: the resulting long reads will subsequently be integrated into the analytical pipeline.





MOBILES

5.4. Metatranscriptomic Sequencing (Illumina NovaSeq, PE150)

Metatranscriptomic sequencing was performed on total RNA extracted from the soil microbial community. The minimum requirements were a concentration of at least 50 ng/μL, a total amount of at least 1 μg and a volume of at least 20 μL. In addition to quantity, the extract had to satisfy purity criteria, expressed as an OD 260/280 ratio between 1.8 and 2.0 and an OD 260/230 ratio between 1.0 and 2.5, which indicate respectively the absence of protein and of organic or chaotropic contaminants. Integrity was assessed as the RNA integrity number (RIN), for which a threshold of at least 6.5 was specified.

This integrity threshold is demanding for soil-derived RNA, which is recovered from a matrix rich in humic substances and exposed to ubiquitous ribonucleases; a RIN of 6.5 is therefore an objective rather than a routinely attainable value. Where the RIN falls below the threshold, providers commonly accept the sample on the basis of the DV200 metric, that is the proportion of fragments longer than 200 nucleotides, which is more informative than the RIN for partially fragmented material and remains compatible with the short insert sizes used in library construction.

The overwhelming majority of the RNA recovered from soil is ribosomal, so that sequencing total RNA directly would spend almost all of the output on transcripts carrying no information on gene expression. Ribosomal RNA was therefore depleted before library construction using hybridisation-based kits designed for bacterial targets, which captured and removed the 16S and 23S rRNA molecules and thereby enriched the messenger RNA fraction. The residual ribosomal fraction was identified in the rRNA-removal step of the MeToReS pipeline that eliminates it computationally.

Libraries were constructed with a directional, or strand-specific, protocol, which preserves the information on which DNA strand the original transcript derived from. This is essential in a prokaryotic metatranscriptome, where genes are densely packed and frequently overlap on opposite strands, and where the strand of origin is often the only way of assigning a read unambiguously to one gene rather than its neighbour. The depleted RNA was fragmented to an insert size of approximately 200–300 bp and converted to cDNA by first-strand synthesis with random priming followed by strand-preserving second-strand synthesis; the product then underwent end repair, A-tailing, ligation of the P5/P7 adapters and index incorporation. Amplification was limited to approximately 10–15 cycles, because over-amplification introduces duplicates and distorts the relative abundances on which the differential-expression analysis depends. The finished library was verified at approximately 300–400 bp and quantified by qPCR or fluorometric assay.

Sequencing was performed on the Illumina NovaSeq platform in paired-end mode with 150 bp per read (PE150), using the same chemistry described for the amplicon libraries but at substantially higher throughput. The recommended output is at least 12 Gb of sequence per sample. For a matrix as complex as soil this represents a moderate depth, and an output in the range of 15–20 Gb per sample is preferable when the analysis relies on de novo assembly, as is the case for the pipeline



MOBILES

described in Section 6. Basecalling and demultiplexing yielded a pair of FASTQ files (R1 and R2) per sample; the specified quality target was a Q30 of at least 85%, with low adapter contamination.

5.5. Comparison of the Sequencing Strategies

Table 2. Comparison of the three sequencing strategies applied within WP3 and of the analytical workflow associated with each of them.

Feature	16S amplicon (Illumina)	16S amplicon (Nanopore)	Metatranscriptomics (Illumina)
Nucleic acid	DNA	DNA	Total RNA
Input requirement	≥ 200 ng total, ≥ 1 ng/μL	as for Illumina 16S	≥ 1 μg total, ≥ 50 ng/μL
Enrichment strategy	PCR amplification of the V3–V4 region	PCR amplification of the near-full-length 16S gene	hybridisation-based depletion of 16S and 23S rRNA
Library type	amplicon, dual-indexed (i5/i7)	amplicon, barcoded	directional (strand-specific) RNA-seq
Platform	Illumina MiSeq / NovaSeq	Oxford Nanopore MinION	Illumina NovaSeq
Read configuration	paired-end 2 × 250 bp	single-molecule long reads spanning the whole amplicon	paired-end 2 × 150 bp
Typical output per sample	50,000–300,000 read pairs	depends on flow cell and multiplexing level	≥ 12 Gb
Quality target	Q30 > 80–85%	depends on chemistry and basecalling model	Q30 ≥ 85%
Primary deliverable	demultiplexed paired FASTQ (R1/R2)	basecalled and demultiplexed FASTQ Dedicated long-read workflow at CNR-ISAFO; not addressed in this deliverable.	demultiplexed paired FASTQ (R1/R2)
Downstream analysis	Section 5 (DADA2 workflow)		Section 6 (MeToReS pipeline)

6. 16S rRNA Amplicon Sequencing for Metataxonomic Analysis

6.1. Overview

Metataxonomic analysis of soil microbial communities was performed by sequencing the V3–V4 hypervariable regions of the bacterial 16S rRNA gene. This region provides sufficient phylogenetic resolution to classify bacteria at the genus level while maintaining a manageable amplicon length (~450 bp) suitable for paired-end Illumina sequencing with 250 bp reads.

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

16





MOBILES

The complete bioinformatic workflow consists of four main stages: (1) preliminary quality assessment of raw reads, (2) adapter and primer trimming, (3) post-trimming quality verification, and (4) denoising and taxonomic classification using the DADA2 pipeline. All scripts are available in the project repository and were executed on the MOBILES HPC infrastructure (hostname: plantilla-gpu0). The complete workflow, from raw reads to the statistical analysis described in Sections 6 and 7, is summarised in Figure 1.



16S rRNA amplicon sequencing — metataxonomic workflow

From raw paired-end reads to ASVs, taxonomy and community statistics

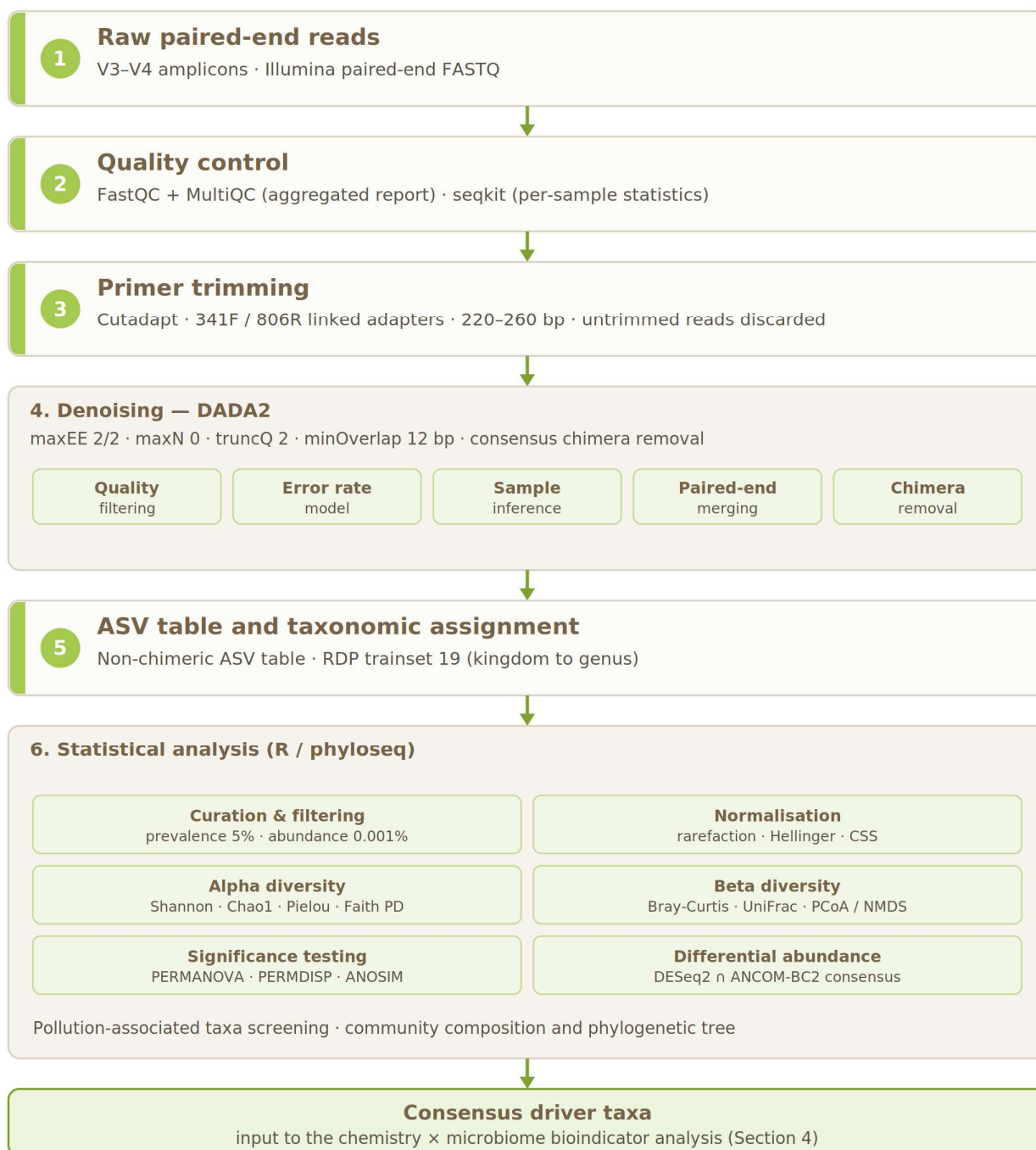


Figure 1. Overview of the 16S rRNA metataxonomic analysis workflow, from raw amplicon reads to community statistics and consensus driver taxa.

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

18



6.2. Preliminary Quality Control

Prior to any processing, raw FASTQ files were subjected to quality assessment at two levels: a global quality control step and a sample-level preliminary analysis.

6.2.1. Global Quality Control (FastQC / MultiQC)

All raw FASTQ files were evaluated using FastQC (v0.11+) to generate per-sample quality reports covering per-base quality scores, sequence length distribution, GC content, adapter content, and duplication levels. MultiQC was then used to aggregate all individual FastQC reports into a single interactive HTML summary, enabling rapid cross-sample comparison. Both tools were run using the following shell script:

```
fastqc -t 4 -o $FQ_OUT $RAW/*.fq.gz && multiqc $FQ_OUT -o $MQC_OUT
```

6.2.2. Sample-level Preliminary Analysis (seqkit)

For each sample, paired forward (*_good_1.fq.gz) and reverse (*_good_2.fq.gz) reads were analysed using seqkit stats to compute per-sample statistics including read count, minimum, mean and maximum read length, and length distribution quartiles. A combined summary TSV table was generated across all samples. Additionally, length distribution profiles were computed by randomly sampling 10,000 reads per file. The analysis was run on all samples in a batch loop, automatically detecting forward-reverse pairs.

6.3. Adapter and Primer Trimming (Cutadapt)

Primer sequences flanking the V3-V4 amplicons (Table 3) were removed using Cutadapt (v3.0+). The primer sequences used correspond to the broadly employed 341F/806R primer pair:

Table 3. Oligonucleotide sequences used in 16S rRNA (V3–V4) amplicon library preparation and metataxonomic read processing.

Oligonucleotide	Stage / role	Sequence (5'→3')
341F— forward primer	V3–V4 amplification / primer removal	ACTCCTACGGGAGGCAGCA
806R— reverse primer	V3–V4 amplification / primer removal	GGACTACHVGGGTWTCTAAT
341F — reverse complement	in-silico primer removal (Cutadapt)	TGCCTGCCTCCCGTAGGAGT
806R — reverse complement	in-silico primer removal (Cutadapt)	ATTAGAWACCCBDGTAGTCC

Degenerate positions follow IUPAC nomenclature (H = A/C/T; V = A/C/G; W = A/T; B = C/G/T; D = A/G/T)

Cutadapt was configured in linked-adapter mode to simultaneously remove the forward primer from the 5' end of R1 (and its reverse complement from the 3' end of R1), and the reverse primer from the 5' end of R2 (with the corresponding adapters on R2). Only reads in which both primers were





MOBILES

successfully detected and removed were retained (`--discard-untrimmed`). Read length was filtered to retain only fragments between 220 and 260 bp per read, consistent with the expected V3-V4 amplicon structure after primer removal. Reads containing ambiguous bases (N) were discarded (`--max-n 0`). No additional quality trimming was applied at this stage, as quality filtering is performed within DADA2.

The trimming was run in a batch loop over all sample pairs, and a log file was saved per sample for traceability.

6.4. Post-trimming Quality Control

After primer trimming, a post-trimming quality check was performed for all samples using `seqkit stats`, reporting read counts and length statistics (min, mean, max) for both R1 and R2 trimmed files. In addition, residual primer contamination was assessed by searching for the forward primer sequence and its reverse complement directly within the trimmed read files using `zgrep` pattern matching. Results were compiled into a per-sample TSV report (`post_trim_summary.tsv`) to verify that trimming was successful and that no significant primer carry-over remained.

6.5. DADA2 Pipeline: Denoising and ASV Generation

The core metataxonomic analysis was performed using the DADA2 R package (v1.28+), which implements a probabilistic error-correction model to resolve exact Amplicon Sequence Variants (ASVs) from Illumina amplicon data. The pipeline was executed via a dedicated R script (`dada2_standard_RDP.R`) using standard DADA2 default parameters, run under R version 4.5.2 on the MOBILES HPC node.

6.5.1. Quality Filtering

Reads were quality-filtered using the `filterAndTrim()` function with the following parameters: maximum expected errors per read (`maxEE`) set to 2 for both forward and reverse reads, no reads with ambiguous bases (`maxN = 0`), truncation at quality score Q2 (`truncQ = 2`), and no additional truncation of read length (`truncLen = 0`, reads were already trimmed to optimal length by `Cutadapt`). A minimum read length of 200 bp was enforced. Filtered reads were written to compressed FASTQ files and cached for pipeline resumption.

Prior to filtering, quality plots were generated for up to five representative samples per run using `plotQualityProfile()`, allowing visual inspection of per-base quality trends prior to parameter selection.

6.5.2. Error Rate Learning

DADA2's parametric error model was learned separately for forward and reverse reads using the `learnErrors()` function, which estimates the run-specific error rates from a subset of the reads. The fitted error model was visualised and saved for quality inspection. Warning messages related to log-





MOBILES

scale transformation of zero-count bins in the error plots may be observed but are expected and do not affect the analysis.

6.5.3. Sample Inference (Denoising)

For each sample, unique sequences were dereplicated using `derepFastq()` and then processed by the DADA2 core sample inference algorithm (`dada()`) using the learned error model. Denoising was performed independently for forward and reverse reads.

6.5.4. Paired-end Merging

Denoised forward and reverse reads were merged using `mergePairs()` with a minimum overlap of 12 bp (DADA2 default) and zero mismatches allowed in the overlap region (`maxMismatch = 0`). Prior to merging, a read-length check is performed to verify that the forward and reverse reads provide sufficient overlap for the expected V3-V4 amplicon length, computed as (forward length + reverse length - 450 bp).

6.5.5. Sequence Table Construction and Chimera Removal

Merged reads from all samples were assembled into a sequence table (`makeSequenceTable()`). Chimeric sequences were then removed using `removeBimeraDenovo()` in consensus mode. The resulting non-chimeric sequence table and an ASV length-distribution histogram were saved to the output directory.

6.5.6. Taxonomic Assignment

Taxonomic classification of ASVs was performed using the `assignTaxonomy()` function with the RDP trainset 19 reference database (`rdp_19_toGenus_trainset.fa.gz`), which provides classification from Kingdom to Genus level. A read-tracking table recording the number of reads retained at each stage of the pipeline is generated to support per-run quality assessment.

6.6. Pipeline Summary and Outputs

The key parameters governing the 16S rRNA metataxonomic workflow, from primer trimming through to taxonomic assignment, are summarised in Table 4. The main output files generated by the pipeline are listed at the end of this section.

Table 4. Key parameters of the 16S rRNA metataxonomic pipeline

Parameter	Value
Primer pair (V3–V4)	341F / 806R
Read-length filter (Cutadapt)	220–260 bp per read
maxEE (forward / reverse)	2 / 2
maxN	0
truncQ	2
truncLen	0 (reads already primer-trimmed)
minOverlap	12 bp (DADA2 default)
maxMismatch	0 (DADA2 default)
Chimera removal	<code>removeBimeraDenovo</code> (consensus mode)
Taxonomy database	RDP trainset 19 (to genus level)





MOBILES

Output files saved by the pipeline include: the non-chimeric sequence table (sequence_table_nochim.rds), taxonomic assignments (tax_assignment.rds), read tracking summary (read_tracking_summary.txt), merge statistics (merge_statistics.txt), and quality visualisation plots (quality_plots/).

6.7. Statistical Analysis of the Metataxonomic Data

The ASV tables produced by the workflow described above were analysed with a dedicated statistical pipeline implemented in R, built on the phyloseq framework. The same pipeline was applied to the two MOBILES sampling campaigns, spring 2025 and autumn 2025, with identical processing steps, parameters and statistical tests; the two seasons differ only in the input dataset and in the sample-to-site correspondence. The analysis is described here once and applies to both campaigns. All steps use a fixed random seed so that rarefaction and permutation-based tests are exactly reproducible.

6.7.1. Data Curation and Filtering

The ASV table, the taxonomic assignments, the phylogenetic tree and the sample metadata (country of origin, contamination status and sampling campaign) are combined into a single phyloseq object. The taxonomy is then curated: ASVs with no counts are discarded, only Bacteria and Archaea are retained, and ASVs unassigned at phylum level are removed. Two successive filters are applied to reduce the sparsity typical of amplicon data: a prevalence filter, which discards ASVs present in fewer than 5% of the samples, and an abundance filter, which discards ASVs accounting for less than 0.001% of the total reads. Sequencing depth is inspected before and after filtering, and the number of ASVs retained at each step is reported.

6.7.2. Normalisation

Because different analyses are sensitive to different aspects of the count data, three normalisation strategies are prepared in parallel from the filtered object. Alpha diversity is computed on a rarefied table, obtained by subsampling all samples to the depth of the smallest one, which is the standard approach when comparing richness estimators across samples of unequal depth. Beta diversity is computed on a Hellinger-transformed table, which down-weights the influence of the most abundant taxa on the distance measures. A cumulative sum scaling (CSS) normalised object is also generated and kept available for analyses that require normalised counts, such as co-occurrence networks or external tools.

6.7.3. Alpha Diversity

Within-sample diversity is characterised with complementary indices: observed richness, Chao1, the Shannon index, Pielou's evenness and Faith's phylogenetic diversity, the latter computed on the midpoint-rooted phylogenetic tree so that the evolutionary breadth of each community is taken into account. Differences between groups are assessed with a non-parametric Kruskal-Wallis test,





MOBILES

followed by Dunn's post-hoc test with Bonferroni correction and by pairwise Wilcoxon tests with Benjamini-Hochberg correction; a parametric linear model is fitted in parallel as a cross-check. Results are reported as boxplots by group together with the corresponding test statistics.

6.7.4. Beta Diversity

Between-sample community structure is explored with two families of distances. Bray-Curtis dissimilarities, computed on the Hellinger-transformed table, describe differences in composition and relative abundance; weighted and unweighted UniFrac distances, computed on the filtered table, additionally incorporate the phylogenetic relatedness of the taxa, with the unweighted variant emphasising the presence or absence of lineages and the weighted variant their abundance. Each distance matrix is ordinated by Principal Coordinates Analysis (PCoA) and, for Bray-Curtis, also by non-metric multidimensional scaling (NMDS) with the associated stress value; ordinations are annotated by country and by contamination status.

The significance of the grouping is tested with PERMANOVA, complemented by pairwise PERMANOVA between groups and by ANOSIM. Because PERMANOVA can return a significant result when groups differ in dispersion rather than in location, a test of multivariate homogeneity of dispersions (PERMDISP) is always run alongside it, and the two results are interpreted together: a significant PERMANOVA accompanied by a non-significant PERMDISP supports a genuine difference in community composition.

6.7.5. Differential Abundance

Taxa differing between groups are identified at genus level, after agglomerating the ASVs by genus, retaining genera present in at least 10% of the samples and discarding samples with an insufficient library size. The uncontaminated reference site is set as the reference level, so that every comparison is expressed as a contrast against the control condition.

Two independent methods are applied to the same data. DESeq2 models the counts with a negative binomial distribution and is comparatively liberal, recovering many true signals but also a tail of false positives arising from the compositional nature of amplicon data. ANCOM-BC2 explicitly estimates and corrects the compositional bias of metabarcoding and is considerably more conservative, additionally reporting a sensitivity analysis that flags results dependent on the treatment of zero counts. Both are run with the same prevalence and library-size settings and with false-discovery-rate correction.

The two result sets are then intersected to obtain a consensus list of driver taxa, retaining only genera declared significant by both frameworks with a consistent direction of change and an effect size above the chosen threshold. This consensus is deliberately restrictive: it represents the taxa whose differential abundance is robust to the statistical framework employed, and it is the set carried forward to the bioindicator analysis described in Section 7.





MOBILES

6.7.6. Pollution-associated Taxa and Community Composition

Independently of the statistical contrasts, the curated dataset is screened against a curated list of genera reported in the literature as associated with specific contamination processes, grouped into four categories: heavy-metal tolerance and transformation, antibiotic resistance, hydrocarbon degradation and plastic degradation. This targeted screening provides a hypothesis-driven complement to the untargeted differential-abundance analysis and highlights taxa of direct interest for the objectives of MOBILES.

Finally, the composition of the communities is summarised as stacked bar charts at the main taxonomic ranks, and the phylogenetic relationships of the dominant phyla are displayed as an annotated circular tree, restricted to the most represented phyla.

7. MeToReS: an End-to-End Snakemake Pipeline for Metatranscriptomic Analysis of Toxicity and Antimicrobial Resistance

7.1. Overview

Metatranscriptomic analysis enables the characterisation of the actively transcribed portion of the soil microbiome, providing information on microbial gene expression and functional activity under different environmental conditions or across different soil compartments. Within the MOBILES project, total RNA extracted from soil samples (as described in D3.1) is subjected to ribosomal RNA depletion prior to sequencing, yielding messenger RNA-enriched libraries that are processed through the MeToReS pipeline.

MeToReS (an end-to-end Snakemake pipeline for metatranscriptomic analysis of toxicity and antimicrobial resistance) is a Snakemake-based workflow that automates the complete analysis from raw paired-end FASTQ files to differential-expression results, functional annotations, community-composition profiles, constrained (canonical) ordination, resistance-gene profiling, and integration with the physico-chemical characterisation of the samples, together with a comprehensive set of downstream visualisations. The pipeline runs within a Conda environment, ensuring reproducibility and consistent software dependencies across runs, and can be executed either on a single machine or submitted to an HPC scheduler (SLURM). The grouping of samples into experimental conditions and the reference level for the comparisons are defined in the configuration file, so the same pipeline can be applied to different experimental designs without modification. It was developed collaboratively by CNR-ISAFOM and UPNA within the MOBILES WP3 framework.





7.2. Pipeline Workflow

The analysis proceeds through a series of interconnected stages, summarised in Table 5 and illustrated in Figure 2. The workflow is defined in a central Snakefile, which orchestrates all analysis steps as a series of interconnected rules. Each rule specifies its inputs, outputs, tools, and parameters, allowing Snakemake to automatically determine execution order, manage inter-step dependencies, and parallelise tasks where possible. The Snakefile is configured via two user-provided files:

- **Sample sheet (CSV):** contains sample names, the grouping variable(s) used to define the experimental conditions, and optionally SRA accession numbers for automated download.
- **Configuration file (YAML):** specifies all analysis parameters, including the grouping column and reference level, QC settings, assembly thresholds, differential-expression cut-offs, taxonomic-classification parameters, functional-annotation database paths, and the optional constrained-ordination, resistance-annotation and chemical-integration modules.

Input data can be provided either as local FASTQ files placed in the designated input directory (results/fastq/) or downloaded automatically from the NCBI Sequence Read Archive (SRA) via SRA accession numbers in the sample sheet.

To launch the pipeline, users activate the dedicated Conda environment (conda activate MeToReS) and run Snakemake. On a single machine the workflow is executed with a local command specifying the number of CPU cores; on an HPC system it is submitted to a SLURM scheduler through the provided cluster profile (cluster.yaml), which maps each rule to an sbatch submission with its own memory, CPU and wall-time requests. Adequate memory must be available for the Kaiju classification step, which loads the full ProGenomes database index (80-100 GB RAM):

```
snakemake --cores 32 --resources mem_mb=110000
```

Table 5. Summary of the pipeline stages in the MeToReS metatranscriptomics workflow

Pipeline Stage	Description
1. Raw samples	Paired-end FASTQ files provided locally or downloaded via SRA accession numbers
2. Quality control	FastQC + MultiQC (initial QC), Trimmomatic (adapter trimming and quality filtering), SortMeRNA (rRNA removal)
3. Per-sample assembly	MEGAHIT: de novo assembly of each sample independently; Prodigal ORF prediction (CDS and proteins)
4. Gene catalog	Pooling of the per-sample ORFs and redundancy removal with CD-HIT-EST (95% identity)
5. Assembly QC	MetaQUAST (reference-free) and seqkit stats on the pooled catalogue and per-sample assemblies
6. Taxonomic classification	Kaiju against the ProGenomes database (gene catalogue and per-sample reads)





MOBILES

7. Expression quantification	Salmon quasi-mapping (TPM); gene filtering (TPM ≥ 1 in >50% of the samples of a group)
8. Differential expression	DESeq2, multi-contrast (each group vs the reference level); Wald test; BH FDR correction
9. Functional annotation	eggNOG-mapper with DIAMOND: COG, GO, KEGG assignments (all genes and per-contrast DEGs)
10. Downstream analysis	PCoA, ANOSIM, UpSet plots, DEG-count summaries, GO/KEGG enrichment, expression heatmaps, community composition
11. Constrained ordination	db-RDA on VST-normalised counts; partial/conditioned models; permutation ANOVA (optional)
12. Resistance annotation	CARD/RGI (ARGs) and BacMet/DIAMOND (MRGs); resistant-DEG report and expression heatmaps (optional)
13. Chemical data integration	envfit of chemical variables on the expression ordination, chemical PCA, and correlation of ARG/MRG abundances with chemical data (optional)
14. Bioindicator screening	Partial correlation between gene expression and chemical variables, controlling for confounding factors (optional)
15. Multivariable association	MaAsLin2 mixed-effects models associating gene expression with chemical variables, with random effects (optional)



MeToReS — metatranscriptomic (RNA-seq) workflow

Snakemake pipeline: from raw reads to expression, function and integrative modules

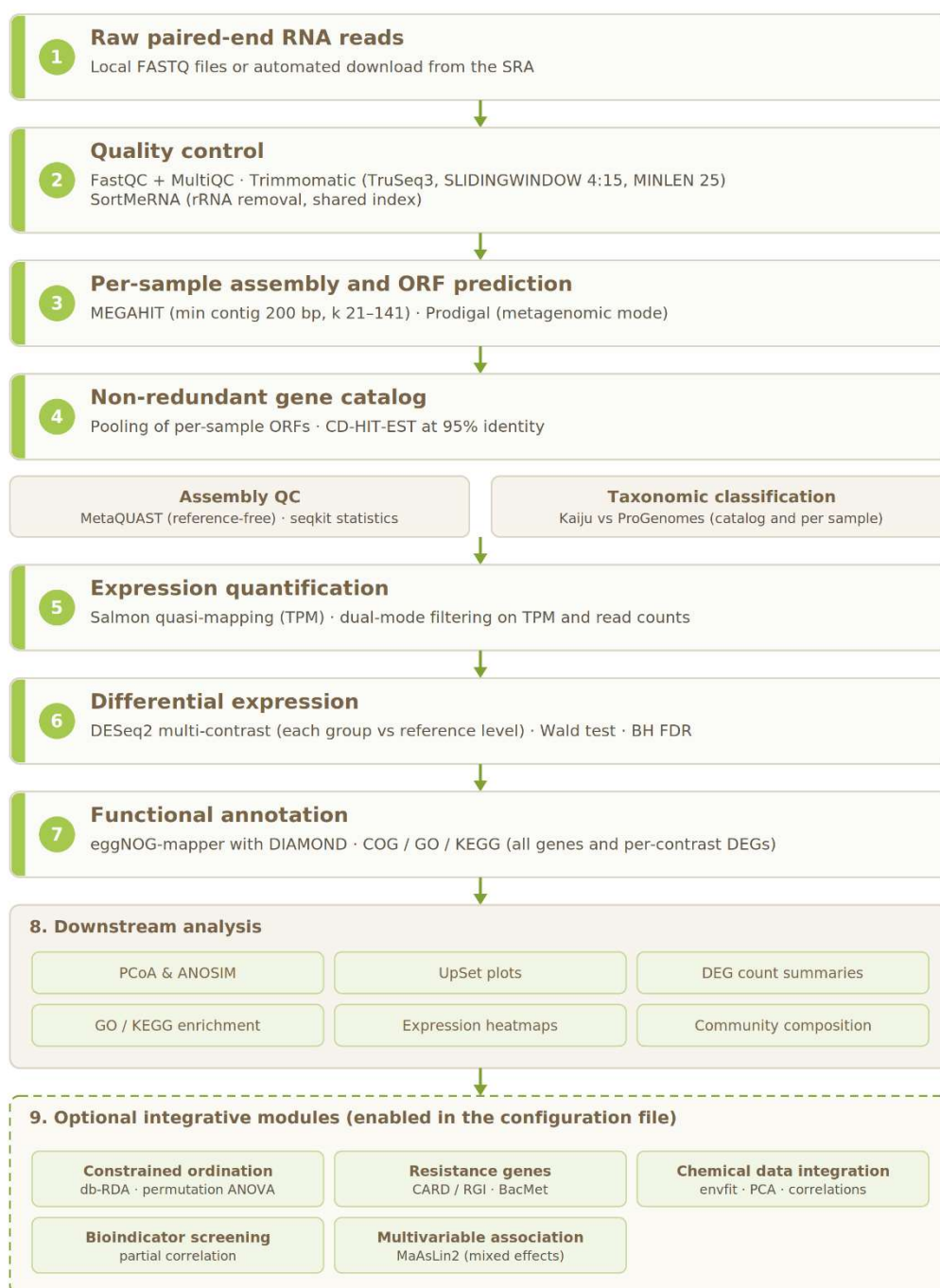


Figure 2. Overview of the MeToReS metatranscriptomic analysis pipeline, from raw RNA reads to differential expression, functional annotation and the optional integrative modules.

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

27



MOBILES

7.3. Quality Control

Quality control is performed in four sequential steps to ensure that only high-quality, informative reads are carried forward into the assembly and quantification stages.

7.3.1. Initial Quality Control

Raw paired-end FASTQ files are evaluated using FastQC to generate per-sample quality reports in HTML format. These reports provide an overview of per-base quality scores, adapter content, GC content, and sequence duplication levels. All FastQC reports are then aggregated into a single interactive summary report using MultiQC, enabling efficient cross-sample comparison prior to trimming.

7.3.2. Adapter Trimming and Quality Filtering

Raw reads are processed using Trimmomatic to remove adapter sequences and low-quality bases. Adapters are clipped using the Illumina TruSeq3 paired-end adapter set (ILLUMINACLIP), leading and trailing low-quality bases are trimmed (LEADING:5, TRAILING:5), and a sliding-window quality filter is applied (SLIDINGWINDOW 4:15); read pairs shorter than the minimum length (MINLEN, 25 bp by default) are discarded. Both paired and unpaired reads are retained as separate output files per sample. The quality cut-off, minimum length and adapter path are all defined in the configuration file.

7.3.3. Ribosomal RNA Removal

Since metatranscriptomic samples are inherently enriched in ribosomal RNA (rRNA), the trimmed reads are screened against curated rRNA reference databases using SortMeRNA, which performs a sensitive sequence-similarity search specifically designed for the depletion of ribosomal transcripts. The rRNA index is built once, in a dedicated step, and is then shared by all samples: this guarantees that the index is created exactly once even when several samples are processed in parallel, and makes the depletion step reproducible across runs. Read pairs that do not match the rRNA references are retained as messenger RNA-enriched reads (rmrRNA FASTQ files) and are carried forward to assembly and quantification, while the alignment log of each sample is stored to document the proportion of ribosomal reads removed. This depletion step is critical to maximise the informative fraction of reads carrying coding-sequence information.

7.4. Assembly and Gene Catalogue Construction

The assembly stage constructs a non-redundant gene catalogue from the cleaned reads. Each sample is assembled independently, open reading frames are predicted per sample, and the resulting genes are pooled and de-duplicated into a single reference catalogue used for all subsequent quantification and annotation steps.





MOBILES

7.4.1. Per-sample Assembly and ORF Prediction

The rRNA-depleted reads of each sample are assembled independently using MEGAHIT, a memory-efficient de novo assembler optimised for metagenomic and metatranscriptomic data (minimum contig length 200 bp; k-mer range 21-141, step 12). Assembling each sample separately avoids the chimeric contigs that can arise when pooling heterogeneous samples and scales efficiently across large sample sets. Open reading frames (ORFs) are then predicted from each sample's contigs using Prodigal in metagenomic mode, producing coding DNA sequences (CDS) and their protein translations; every ORF header is prefixed with the sample identifier to guarantee unique names when the per-sample gene sets are pooled.

7.4.2. Assembly Quality Control

Assembly quality is evaluated using MetaQUAST in reference-free mode, computing standard metrics including total contig count, total assembly size, N50, and GC content, both on the pooled gene catalogue and as per-sample statistics. seqkit stats provides complementary sequence-level statistics, and configurable warning thresholds (minimum N50, total assembly size, and contig count) flag low-quality assemblies. Results are integrated into the final MultiQC report.

7.4.3. Gene Catalogue Construction and Taxonomic Classification

The per-sample ORF sets are concatenated and redundancy is removed using CD-HIT-EST at a 95% sequence-identity threshold, producing a single non-redundant gene catalogue that serves as the reference for all subsequent quantification and annotation steps.

Taxonomic classification of ORFs is performed using Kaiju in catalogue mode, performing protein-level classification by translating each ORF in all six reading frames and searching against the ProGenomes reference protein database. Classification is additionally run per-sample on the rRNA-depleted reads to generate individual taxonomic profiles. The reference database, taxonomic rank for summary plots, and host taxon filter are all configurable in the config.yaml file.

7.5. Gene Expression Quantification

Quantification is performed in five steps to generate a filtered and taxonomy-annotated abundance matrix. A Salmon index is first built from the non-redundant gene catalogue to enable fast quasi-mapping-based quantification. For each sample, rRNA-depleted reads are mapped against the gene index using Salmon, producing transcript-level abundance estimates (TPM — Transcripts Per Million) with bias corrections for GC content and sequence composition.

Individual per-sample quantification files are merged into a single gene expression matrix using a custom Python script. Low-expression genes are then removed by a dual-mode filter that evaluates both the normalised expression (TPM) and the underlying read counts: a gene is retained when it shows consistent expression evidence in a sufficient proportion of the samples of at least one experimental group, so that genes supported only by sporadic or very low counts are discarded. Requiring agreement between the two measures avoids retaining genes whose apparent expression





MOBILES

derives from normalisation artefacts in low-depth samples. Finally, the filtered abundance matrix is annotated with taxonomic information by merging it with the Kaiju gene-level taxonomy table.

7.6. Differential Expression Analysis

Differential expression analysis is performed with DESeq2. Per-sample Salmon quantification files are imported using tximport with length-scaled TPM normalisation, treating each ORF as a gene-level feature, and low-count genes are pre-filtered (minimum total count of 50 across all samples). Samples are grouped according to the grouping column defined in the configuration file, and one contrast is generated automatically for each non-reference group against the configured reference level, so that several experimental conditions are compared within a single run. Differential expression is tested using the Wald test, with p-values corrected for multiple testing by the Benjamini-Hochberg (BH) method to control the false discovery rate (FDR). Genes are classified as up- or down-regulated according to user-defined thresholds (default: $|\log_2FC| > 1$ and adjusted p-value < 0.05). $2FC > 1$, $p_{adj} < 0.05$).

Diagnostic plots are produced to assess the analysis, including a PCA plot computed from VST-normalised counts (showing group separation), a dispersion-estimates plot assessing model fit, and a volcano plot for each contrast. For every contrast, the identifiers, DNA sequences and protein translations of the up- and down-regulated genes are extracted from the reference gene catalogue using seqkit, and an across-contrast summary table is compiled. $2FC$ vs. $-\log_{10}$ adjusted p-value), an MA plot (mean expression vs. fold change), and a dispersion estimates plot assessing model fit. DNA and protein sequences of differentially expressed genes are extracted from the reference gene catalogue using seqkit for downstream functional analyses.

7.7. Functional Annotation

Functional annotation is performed using eggNOG-mapper, which assigns functional information to genes based on sequence homology to the eggNOG ortholog database, using DIAMOND as the search engine. Annotation is carried out at two levels: (1) all filtered genes, providing a comprehensive functional background for the enrichment analyses; and (2) the differentially expressed genes of each contrast, annotated separately for the up- and down-regulated sets to characterise the functions associated with each experimental condition.

The annotations produced include COG (Clusters of Orthologous Groups) functional categories, Gene Ontology (GO) terms, KEGG orthology (KO) identifiers, and KEGG pathway assignments.

7.8. Downstream Analysis

The downstream analysis stage is fully integrated into the Snakemake workflow and executed automatically as part of a standard pipeline run. It integrates expression data with functional annotations to generate biological insights. Key analytical modules include:





MOBILES

- **Host contamination removal:** genes assigned to Streptophyta (plant host) by Kaiju are filtered from the abundance matrix to focus on microbial transcripts.
- **Community-level ordination (PCoA):** Principal Coordinates Analysis on Bray-Curtis dissimilarity matrices of gene expression profiles, with ANOSIM significance testing.
- **Shared and unique gene analysis:** UpSet-plot-based identification of the differentially expressed genes shared between, or unique to, the different contrasts, computed for all DEGs and for the up-regulated set separately.
- **Functional profiling of upregulated genes:** expression heatmaps of the up- and down-regulated genes of each contrast, aggregated by functional category (eggNOG descriptions) and filtered to genes whose combined expression exceeds a configurable threshold.
- **Per-sample community composition:** stacked bar plots of the top 15 most abundant phyla from Kaiju per-sample classification, grouped by experimental condition.
- **GO enrichment analysis:** Gene Ontology over-representation analysis (ORA) on upregulated genes using a hypergeometric test with BH correction, with results visualised as bar plots and dot plots.
- **KEGG pathway enrichment analysis:** KEGG over-representation analysis on upregulated genes, with pathway–gene mappings derived from eggNOG-mapper output.

All downstream outputs are written to a group-specific downstream directory (results/downstream_<group>/) and include ordination plots, ANOSIM results, UpSet plots, per-contrast DEG-count summaries, gene-set files, up- and down-regulated expression heatmaps, COG and GO functional-classification plots, GO/KEGG enrichment plots, and the per-sample community-composition bar plot.

7.9. Constrained Ordination (db-RDA)

In addition to the unconstrained ordination performed during the downstream analysis, the pipeline includes an optional constrained (canonical) ordination step based on distance-based Redundancy Analysis (db-RDA). This analysis quantifies how much of the variation in the gene-expression profiles can be attributed to the experimental metadata variables (for example contaminant, season, and country) and tests their statistical significance.

db-RDA is computed on the VST-normalised expression matrix using a configurable dissimilarity metric (Bray-Curtis by default). Partial (conditioned) models are supported, allowing the effect of one variable to be tested while controlling for another (for example, the contaminant effect conditioned on season, or on country). The significance of each model and of the individual constraints is assessed by permutation-based ANOVA (999 permutations by default), and the proportion of variance explained by each term is reported. The step is enabled and fully parameterised through the configuration file.





7.10. Resistance Gene Annotation (ARG and MRG)

The pipeline includes an optional module for the detection and quantification of resistance determinants in the metatranscriptome, covering both antibiotic resistance genes (ARGs) and metal/biocide resistance genes (MRGs). The non-redundant gene catalogue is first translated into protein sequences (bacterial genetic code 11), preserving the original gene identifiers so that resistance annotations can be cross-referenced with the expression, taxonomy and functional-annotation tables. Annotation is then carried out along two complementary tracks:

1. **Antibiotic resistance (ARGs):** protein sequences are screened against the Comprehensive Antibiotic Resistance Database (CARD) using the Resistance Gene Identifier (RGI) with DIAMOND alignment. Only high-confidence hits (Perfect and Strict; Loose hits discarded) are retained and reported per gene.
2. **Metal and biocide resistance (MRGs):** protein sequences are searched against the BacMet experimentally confirmed database with DIAMOND (blastp), keeping hits above configurable identity and query-coverage thresholds (80% by default) and an e-value cut-off of 1e-10; each hit is mapped to its associated compound (metal or biocide).

The CARD and BacMet annotations are merged with the abundance matrix, the Kaiju taxonomy and the eggNOG functional annotation to produce a unified resistance table together with per-gene abundance summaries aggregated by antibiotic-resistance ontology (ARO) term and by compound. Finally, the resistance genes that overlap the differentially expressed genes of each contrast are tabulated, and expression heatmaps of the top-ranked ARGs and MRGs are generated. Like the constrained-ordination step, this module is optional and controlled entirely through the configuration file.

7.11. Chemical Data Integration

The pipeline can integrate the physico-chemical characterisation of the samples (for example the concentrations of contaminants and trace elements) with the transcriptomic results, in order to relate microbial gene expression and resistance profiles to the measured contamination. Chemical measurements are read from an external spreadsheet and matched to the sequencing samples, and are made available in two fractions, the bioavailable and the pseudo-total fraction, with the bioavailable fraction used by default. The analysis is then carried out along two complementary lines:

3. **Chemical ordination:** the chemical variables are related to the transcriptomic ordination by fitting environmental vectors (envfit) onto the VST-normalised expression matrix, with significance assessed by permutation (999 permutations by default); a complementary principal component analysis (PCA) of the chemical data is also produced.
4. **Resistance-chemistry correlation:** the abundances of the resistance genes, ARGs aggregated by antibiotic-resistance ontology (ARO) term and MRGs aggregated by compound, are correlated with the chemical variables using a configurable correlation





MOBILES

method (Spearman by default), yielding a full correlation matrix, a table of element and compound associations, and a correlation heatmap.

This module links the biological response captured by the metatranscriptome to the environmental chemistry of each site, providing a direct connection between the observed gene-expression and resistance patterns and the contamination measured in the field. Like the other integrative modules, it is optional and controlled entirely through the configuration file.

7.12. Bioindicator Screening by Partial Correlation

Correlating gene expression with contamination directly can be misleading when both vary along a third factor, such as the season or the sampling site: an apparent association may simply reflect that shared structure. To address this, the pipeline provides a partial-correlation module that measures the association between the expression of each gene and each chemical variable after removing the effect of one or more configurable covariates.

The analysis is restricted to the most variable genes of the VST-normalised matrix, in order to concentrate the testing effort on the informative fraction of the transcriptome and to limit the multiple-testing burden. Chemical concentrations are transformed before testing (log10 followed by z-score standardisation by default) so that elements measured on very different scales contribute comparably. Partial correlations are then computed with a configurable method (Spearman by default), p-values are corrected for multiple testing, and gene-element pairs are retained as candidates only when they exceed both a correlation-strength threshold and an adjusted-significance threshold. The module returns the complete set of tested associations together with the filtered list of significant candidates.

7.13. Multivariable Association Modelling (MaAsLin2)

As a complement to the pairwise approach described above, the pipeline integrates MaAsLin2, a framework for the multivariable association of high-dimensional omics features with sample metadata. Rather than testing one variable at a time, MaAsLin2 fits a linear model for each gene in which the chemical variables of interest are entered as fixed effects, optional additional covariates can be included to adjust for confounding, and a grouping factor such as the country of origin is declared as a random effect.

Modelling the sampling site as a random effect is important for a multi-country design such as MOBILES, because samples from the same site are not statistically independent: the mixed-effects formulation absorbs the between-site baseline differences and therefore attributes to the chemical variables only the variation that is genuinely associated with them. Expression data are transformed consistently with the partial-correlation module and the analysis is restricted to the most variable genes. Associations are reported with their effect sizes and adjusted significance, and the significant subset is exported separately as the candidate list. Both this module and the partial-correlation screening are optional and are configured entirely through the configuration file.

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

33





MOBILES

8. Integration of Chemical and Metataxonomic Data for Bioindicator Identification

A central objective of MOBILES is to establish whether the soil microbiome can be used to detect and monitor anthropogenic contamination. This requires moving beyond the description of which taxa differ between sites, and asking whether the abundance of specific taxa tracks the measured concentration of specific contaminants in a way that is strong, stable and reproducible enough to be used as an indicator. This section describes the analytical strategy developed to answer that question by combining the metataxonomic data of Section 6 with the physico-chemical characterisation of the same soils.

The strategy is deliberately conservative, and is organised as a sequence of increasingly stringent filters. Each step is designed to remove a specific class of spurious association, so that only the candidate bioindicators that survive every stage are proposed. The rationale is that, in a design with a limited number of sites and a large number of taxa and elements, statistical significance alone is a weak guarantee: a credible bioindicator must also be biologically coherent, insensitive to individual sites, and reproducible across seasons.

8.1. Chemical Data and Preparation

The chemical characterisation comprises the concentrations of a panel of elements, including both potentially toxic trace elements and major nutrients, measured on the same soils used for sequencing. Concentrations are available as distinct extraction blocks, describing respectively the fraction that is readily available to organisms, the pseudo-total content, and the ratio between the two, which expresses the mobile proportion of each element. The three blocks are analysed in separate models rather than jointly, because they are strongly interdependent by construction and because they answer different questions: the available fraction is the one to which microorganisms are actually exposed, whereas the pseudo-total content describes the overall contamination load of the site.

Values below the analytical detection limit are imputed as half of the smallest positive value observed for that element, a standard convention that preserves the ranking of the samples without introducing artificial zeros. Elements that are entirely below the detection limit or constant across the dataset are discarded. The remaining concentrations are log-transformed and then standardised to zero mean and unit variance, so that elements measured on different scales contribute comparably to the multivariate models. Chemical measurements and sequencing samples are matched through the sampling site, and the correspondence is verified and reported before any analysis is run.

8.2. Community-level Relationship with the Chemistry

The first question addressed is whether the chemistry explains a significant share of the variation in community composition at all. This is tested with a transformation-based redundancy analysis (tb-





MOBILES

RDA), in which the Hellinger-transformed community table is constrained by the chemical variables. Because the number of elements is large relative to the number of samples, the explanatory variables are not all entered at once: a forward selection procedure adds elements one at a time, retaining only those that significantly improve the model, and the collinearity of the selected set is checked through variance inflation factors.

The significance of the resulting model, and of each selected element, is assessed by permutation (999 permutations), and collinearity among the selected variables is monitored through variance inflation factors. Permutations are restricted so that samples are exchanged only within their own site: this is an essential precaution, because replicates from the same site share both a similar community and a similar chemistry, and an unrestricted permutation would treat them as independent observations and strongly overstate the significance of the result. The adjusted proportion of variance explained is reported for each model, and the same permutation scheme and number of permutations are used for the Mantel test and the distance-based redundancy analysis.

The same question is addressed from a complementary angle by relating the matrix of community dissimilarities directly to the matrix of chemical distances, through a Mantel test and a distance-based redundancy analysis. Agreement between the constrained ordination and the distance-based approach indicates that the association between chemistry and community structure does not depend on the particular ordination method chosen.

8.3. Taxon-Element Correlations

Having established the community-level relationship, the analysis moves to individual taxa. Abundances are first expressed as centred log-ratios, a transformation appropriate for compositional data such as amplicon counts, which removes the spurious negative correlations that arise when abundances are constrained to sum to a constant. Only taxa present in at least a quarter of the samples are tested, since rare taxa cannot support a reliable correlation.

Each taxon is then correlated with each element using a rank-based (Spearman) coefficient, which does not assume a linear relationship and is robust to the outlying concentrations common in contaminated soils. Only taxa present in at least 25% of the samples are tested, since rare taxa cannot support a reliable correlation. Given the large number of taxon-element pairs, p-values are corrected with the Benjamini-Hochberg procedure and associations are retained only when the adjusted p-value is below 0.05. This step is performed at both ASV and genus level.

8.4. Consensus-anchored, Site-level Screening

The correlations described above are computed across all individual samples, and are therefore vulnerable to two problems: the very large number of tests, and the fact that replicates from the same site are not independent. The decisive filtering step addresses both by restricting the analysis to a





MOBILES

small set of taxa selected a priori, and by evaluating the relationship at the level of sites rather than individual samples.

The candidate taxa are the consensus driver taxa identified in Section 6, that is the genera declared differentially abundant by both DESeq2 and ANCOM-BC2. Anchoring the analysis to this consensus is what makes the screening credible: these are the taxa whose response to the contamination gradient is already supported by two independent statistical frameworks, so testing only them both reduces the multiplicity of the analysis by orders of magnitude and ensures that any association found concerns a taxon with an established biological signal.

For each candidate taxon-element pair, the abundances and the concentrations are averaged within each site and correlated across sites, giving one observation per site. Working on site means removes the pseudo-replication of the within-site samples and asks the question that matters for a monitoring application: does this taxon track the contamination across independent locations? A pair is considered strong when the absolute site-level Spearman coefficient reaches 0.714, which is the critical value for the number of sites available in MOBILES at a significance level of 0.05, and pairs supported by fewer than five sites are not evaluated. Each pair is then subjected to three further checks:

5. **Stability (jackknife):** the site-level correlation is recomputed leaving out one site at a time, and the weakest of these recomputed values is retained. A pair is considered stable when this worst-case absolute coefficient remains at least 0.60. If the association collapses when a single site is removed, it is driven by that site alone rather than by a general gradient, and the pair is rejected.
6. **Directional coherence:** the sign of the correlation is compared with the direction of the differential-abundance result. A taxon enriched in contaminated sites is expected to correlate positively with the corresponding element; a pair whose correlation contradicts the differential-abundance result is not biologically consistent, however strong it may be.
7. **Specificity and within-site consistency:** it is recorded whether the element is among the primary contaminants of the sites in which the taxon is enriched, and whether the sign of the relationship also holds within sites. These are reported as descriptive diagnostics rather than as exclusion criteria.

On the basis of these checks, each pair is assigned to one of four tiers (Table 5), so that the strength of the evidence supporting each candidate is made explicit rather than reduced to a binary significant or non-significant verdict.



Table 6. Classification of the candidate taxon-element pairs by strength of evidence.

Tier	Criteria
A - primary and coherent	Strong and jackknife-stable site-level correlation, element among the primary contaminants of the site, and direction coherent with the differential-abundance result
B - coherent	Strong and jackknife-stable correlation with a direction coherent with the differential-abundance result, for a non-primary element
C - robust only	Strong and jackknife-stable correlation, but without coherence with the differential-abundance result
D - weak	Correlation not reaching the significance threshold or not stable to the jackknife

8.5. Leverage Diagnostics

Because the site-level correlations are computed on a small number of points, a single site with an extreme concentration can dominate the result. For every candidate pair, a diagnostic plot is therefore produced showing the individual samples coloured by site, the site means on which the correlation is actually based, and two trend lines, one fitted to the individual samples and one to the site means. A marked divergence between the two lines is an immediate visual indication that the relationship is driven by leverage rather than by a consistent gradient. Each pair is annotated with its site-level coefficient, its jackknife stability, and a resulting classification ranging from robust to leverage-driven, so that the candidates can be inspected individually before any is proposed as an indicator.

8.6. Cross-season Reproducibility

The final and most stringent filter is the comparison of the two sampling campaigns. The candidate lists obtained independently for spring and autumn are joined without any prior filtering, so that pairs that are strong in one season and absent in the other remain visible: these non-replications are as informative as the confirmations, since they identify relationships that cannot be generalised beyond a single sampling occasion.

Pairs that are strong in both seasons, applying the same site-level and jackknife thresholds used within each campaign, with the same sign and a direction coherent with the differential-abundance results, constitute the reproducible and coherent set, and are the only ones proposed as candidate bioindicators.

The outcome of this procedure is a short, explicitly qualified list of taxon-element associations that are statistically significant, robust to the removal of individual sites, biologically coherent with the differential-abundance analysis, and reproducible across two independent sampling campaigns. This list constitutes the set of candidate bioindicators proposed by WP3, together with the full documentation of the associations that were tested and rejected at each stage.



9. Conclusion

Deliverable 3.2 established the reference protocols and analytical workflows adopted in WP3 for the sequencing-based characterisation of the MOBILES soil microbial communities. For each nucleic acid it specified the acceptance criteria inherited from D3.1, the library preparation, the sequencing configuration and the primary data processing, across three complementary strategies: short-read 16S rRNA amplicon sequencing on the Illumina platform, long-read 16S rRNA amplicon sequencing on the Oxford Nanopore MinION, and metatranscriptomic sequencing on Illumina. It further defined the two reference analytical workflows applied to these data — a DADA2-based metataxonomic workflow for the amplicon data and the MeToReS Snakemake pipeline for the metatranscriptomic data — together with the integration of measured environmental chemistry for the identification of candidate biological indicators.

By fixing these procedures, parameters and workflows in advance, the deliverable ensures that samples processed at different times and by different partners are treated consistently and that the results obtained across the MOBILES sites remain comparable. The reference-free, replicate-aware design of the metatranscriptomic pipeline, the taxonomic and functional annotation of the actively expressed community, and the coupling of expression and resistome profiles with real soil chemistry together provide a reproducible basis for relating microbial activity to contamination rather than to categorical exposure groups alone.

The protocols and workflows described here were applied to the soil samples collected during the MOBILES Spring 2025 campaign. The sequence data they produce are stored and processed on the data and the computational infrastructure is described in the companion deliverable D3.3, while the biological results derived from them — community composition, differentially expressed genes and putative gene clusters, and the microbial taxa and genes proposed as biological indicators — will be reported in deliverable D3.4. Because the analytical workflows are modular and configuration-driven, they can accommodate additional samples, alternative experimental variables or new analytical modules without altering their core structure, so that the reference established here remains applicable as the project's datasets grow.

Within the broader aim of MOBILES — the development of biosensors and monitoring tools for chemical pollutants and antimicrobial resistance in the environment — the protocols and pipelines consolidated in this deliverable provide a standardised, reproducible route from soil samples to the microbial taxa and genes that can serve as biological indicators of contamination. D3.2 therefore furnishes the methodological foundation on which the subsequent identification of these indicators, and their eventual exploitation within the project, will be built.





10. List of Figures

Figure 1. Overview of the 16S rRNA metataxonomic analysis workflow.

Figure 2. Overview of the MeToReS metatranscriptomic analysis pipeline.

11. List of Tables

Table 1. MOBILES WPs list

Table 2. Comparison of the three sequencing strategies applied within WP3 and of the analytical workflow associated with each of them.

Table 3. Oligonucleotide sequences used in amplicon library preparation and metataxonomic read processing.

Table 4. Key parameters of the 16S rRNA metataxonomic pipeline.

Table 5. Summary of the pipeline stages in the MeToReS metatranscriptomics workflow.

Table 6. Classification of the candidate taxon-element pairs by strength of evidence.

12. List of Abbreviations

Abbreviation	Abbreviation for
AMR	Antimicrobial-Resistant
ANCOM-BC2	Analysis of Compositions of Microbiomes with Bias Correction 2
ANOSIM	Analysis of Similarities
ARG	Antibiotic (Antimicrobial) Resistance Gene
ARO	Antibiotic Resistance Ontology
ASV	Amplicon Sequence Variant
CARD	Comprehensive Antibiotic Resistance Database
CDS	Coding DNA Sequence
CECs	Contaminants of Emerging Concern
CLR	Centred Log-Ratio
COG	Clusters of Orthologous Groups
CSS	Cumulative Sum Scaling



DADA2	Divisive Amplicon Denoising Algorithm 2
db-RDA	Distance-based Redundancy Analysis
DEG	Differentially Expressed Gene
DESeq2	Differential Expression Analysis using a Negative Binomial model
EPs	Experimental Places
FDR	False Discovery Rate
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOD	Limit of Detection
MaAsLin2	Multivariable Association Discovery in Population-scale Meta-omics Studies 2
MRG	Metal/Biocide Resistance Gene
NMDS	Non-metric Multidimensional Scaling
ORF	Open Reading Frame
OTU	Operational Taxonomic Unit
PCA	Principal Component Analysis
PCoA	Principal Coordinates Analysis
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
PERMDISP	Permutational Analysis of Multivariate Dispersions
PMPs	Persistent and Mobile Pollutants
QC	Quality Control
RDP	Ribosomal Database Project
RGI	Resistance Gene Identifier
rRNA	Ribosomal RNA
tb-RDA	Transformation-based Redundancy Analysis
TPM	Transcripts Per Million
VIF	Variance Inflation Factor
VST	Variance Stabilising Transformation
WP	Work Package



MOBILES

13. 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/

