Check for updates

OPINION ARTICLE

Establishing the ELIXIR Microbiome Community [version 1; peer review: awaiting peer review]

Robert D. Finn [1], Bachir Balech [2], Josephine Burgin[1], Physilia Chua [3], Erwan Corre[4], Cymon J. Cox[5], Claudio Donati[6], Vitor Martins dos Santos[7], Bruno Fosso[8], John Hancock[9], Katharina F. Heil [3], Naveed Ishaque [10], Varsha Kale[1], Benoit J. Kunath[11], Claudine Médigue[12,13], Evangelos Pafilis[14], Graziano Pesole [2,8], Lorna Richardson [1], Monica Santamaria[15], Tim Van Den Bossche [16,17], Juan Antonio Vizcaíno [1], Haris Zafeiropoulos [14], Nils P. Willassen[18], Eric Pelletier [13,19], Bérénice Batut [12,20]

[1]European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, UK
[2]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Bari, Italy
[3]ELIXIR Hub, Hixton, UK
[4]Station Biologique de Roscoff, CNRS/Sorbonne Universite, Roscoff, France
[5]Centro de Ciências do Mar, Universidade do Algarve, Faro, Portugal
[6]Edmund Mach Foundation Research and Innovation Centre, San Michele all'Adige, Trentino-South Tyrol, Italy
[7]Systems and Synthetic Biology, Wageningen University & Research, Wageningen, Gelderland, The Netherlands
[8]Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Bari, Italy
[9]Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
[10]Berlin Institute of Health Charité, Universitätsmedizin Berlin, Berlin, Germany
[11]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[12]Institut Français de Bioinformatique, CNRS, Evry, France
[13]Genomics Metabolics, Genoscope, Institut François-Jacob / CEA / CNRS / Université Evry / Université Paris-Saclay, Evry, France
[14]Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Greece
[15]Department of Soil, Plant and Food Sciences (Di.S.S.P.A.), University of Bari, Bari, Italy
[16]VIB, UGent Center for Medical Biotechnology, Ghent, Belgium
[17]Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent, Belgium
[18]UiT The Arctic University of Norway, Tromsø, Norway
[19]Research Federation for the study of Global Ocean Systems Ecology and Evolution, Paris, France
[20]Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Freiburg, Germany

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

Abstract

Microbiome research has grown substantially over the past decade in terms of the range of biomes sampled, identified taxa, and the volume of data derived from the samples. In particular, experimental approaches such as metagenomics, metabarcoding, metatranscriptomics and metaproteomics have provided profound insights into the vast, hitherto unknown, microbial biodiversity. The

ELIXIR Marine Metagenomics Community, initiated amongst researchers focusing on marine microbiomes, has concentrated on promoting standards around microbiome-derived sequence analysis, as well as understanding the gaps in methods and reference databases, and solutions to computational overheads of performing such analyses. Nevertheless, the methods used and the challenges faced are not confined to marine studies, but are broadly applicable to all other biomes. Thus, expanding this Community to a more inclusive ELIXIR Microbiome Community will enable it to encompass a broad range of biomes and link expertise across 'omics technologies. Furthermore, engaging with a large number of researchers will improve the efficiency and sustainability of bioinformatics infrastructure and resources for microbiome research (standards, data, tools, workflows, training), which will enable a deeper understanding of the function and taxonomic composition of the different microbial communities.

Keywords
Microbiome, ELIXIR Community, White Paper

This article is included in the ELIXIR gateway.

Corresponding authors: Robert D. Finn (rdf@ebi.ac.uk), Eric Pelletier (eric.pelletier@genoscope.fr), Bérénice Batut (berenice.batut@gmail.com)

Author roles: Finn RD: Writing – Original Draft Preparation, Writing – Review & Editing; **Balech B**: Writing – Original Draft Preparation, Writing – Review & Editing; **Burgin J**: Writing – Original Draft Preparation, Writing – Review & Editing; **Chua P**: Writing – Original Draft Preparation, Writing – Review & Editing; **Corre E**: Writing – Original Draft Preparation, Writing – Review & Editing; **Cox CJ**: Writing – Original Draft Preparation, Writing – Review & Editing; **Donati C**: Writing – Original Draft Preparation, Writing – Review & Editing; **dos Santos VM**: Writing – Original Draft Preparation, Writing – Review & Editing; **Fosso B**: Writing – Original Draft Preparation, Writing – Review & Editing; **Hancock J**: Writing – Original Draft Preparation, Writing – Review & Editing; **Heil KF**: Writing – Original Draft Preparation, Writing – Review & Editing; **Ishaque N**: Writing – Original Draft Preparation, Writing – Review & Editing; **Kale V**: Writing – Original Draft Preparation, Writing – Review & Editing; **Kunath BJ**: Writing – Original Draft Preparation, Writing – Review & Editing; **Médigue C**: Writing – Original Draft Preparation, Writing – Review & Editing; **Pafilis E**: Writing – Original Draft Preparation, Writing – Review & Editing; **Pesole G**: Writing – Original Draft Preparation, Writing – Review & Editing; **Richardson L**: Writing – Original Draft Preparation, Writing – Review & Editing; **Santamaria M**: Writing – Original Draft Preparation, Writing – Review & Editing; **Van Den Bossche T**: Writing – Original Draft Preparation, Writing – Review & Editing; **Vizcaíno JA**: Writing – Original Draft Preparation, Writing – Review & Editing; **Willassen NP**: Writing – Original Draft Preparation, Writing – Review & Editing; **Zafeiropoulos H**: Writing – Original Draft Preparation, Writing – Review & Editing; **Pelletier E**: Writing – Original Draft Preparation, Writing – Review & Editing; **Batut B**: Writing – Original Draft Preparation, Writing – Review & Editing

Introduction

The term "microbiome" is a description of an entire habitat that encompasses all the microbes (bacteria, archaea, eukaryotes, and viruses), their genomes, and the environment they are found in Ref. [1]. The microbiome is experimentally characterised by the application of one or more 'omics techniques, especially metabarcoding, metagenomics and metatranscriptomics, but also metaproteomics and metabolomics, combined with contextual metadata about the surrounding environment, be it a geographic location (e.g. ocean), host-associated (e.g. human gut) or engineered (e.g. wastewater treatment plant). Over the past decade, scientists have become increasingly aware of the role performed by microbes in the health (or maintenance) of the environment, and that dysbiosis of the microbial community can lead to dysregulation and/or negative outcomes. Microbial communities can be very diverse and heterogeneous in composition across geospatial and temporal scales, and the culture-independent methods for identifying species with the microbiome often reveal hitherto unknown microbes. Despite methodological difficulties, understanding the taxonomic and functional composition of a microbiome, how compositional differences relate to phenotypes, and how these communities may be manipulated to restore a community to a natural or normal composition are key, current research questions. Due to the diminishing costs of nucleic acid sequencing and high-availability of sequencing platforms there are now millions of microbiome-derived sequence datasets, many of which are large (gigabytes to terabytes) and complex (thousands of related samples). Additionally, datasets from other 'omics techniques such as metaproteomics are being increasingly generated, alone or in combination with metagenomics and/or metatranscriptomics data coming from the same samples. A key challenge facing the microbiome research community is how to: appropriately store the data; informatically process, integrate, compare and interpret microbiome-derived data; and how to make the data findable, accessible, interoperable and reproducible, i.e. FAIR.[2]

ELIXIR[3] is a distributed infrastructure bringing together experts from across Europe to enable life science researchers throughout the world to access and analyse life science data. ELIXIR is formed by member states each with a national Node composed of one or more centres of excellence in bioinformatics. Each Node coordinates services, standards and resources, and collaborates with experts in other Nodes to create a sustainable Europe-wide infrastructure for biological data. ELIXIR Platforms bring together experts from Nodes to develop ELIXIR's vision and coordinate activities in defined areas. The five Platforms are Data, Tools, Interoperability, Compute and Training. ELIXIR Communities bring together experts across ELIXIR Nodes and external partners to coordinate activities within specific life science domains. The ELIXIR Marine Metagenomics Community acted as a biome-specific network of researchers for the identification and organisation of domain-specific reference resources, development of reproducible workflows and the proposal of best practices. However, there is no underlying reason to restrict these activities to just the marine environment, with most of the aforementioned efforts broadly applicable to analysis of microbiome-derived sequence data from any environment. Furthermore, there is the need to extend the activities of the Community to integrate expertise and knowledge about other 'omics technologies, such as metaproteomics and metabolomics, which are increasingly used in microbiome studies. Thus, this white paper outlines some of the historical aspects of the Community and the aims of the broader community, especially in the context of the other ELIXIR Communities and infrastructure platforms.

From marine metagenomics to a more inclusive Community

The ELIXIR Marine Metagenomics Community, established in 2015 as part of the European Commission funded ELIXIR EXCELERATE project (grant number 676559), was one of the first four ELIXIR Communities created as "Use Cases".[4] During the EXCELERATE project, these ELIXIR "Use Cases" were expanded and renamed to Communities, with a unified aim of bringing European specialists together to provide sustainable data resources, benchmark different tools and workflows, provide access to computing and storage, improve interoperability, and develop training resources within their research domains. These activities were conducted in collaboration with the ELIXIR Platforms, to ensure harmonisation of the outputs. As such, the Marine Metagenomics Community focused on metagenomics analysis pipelines, addressing the lack of reference databases and promoting the best practices for the research community. Highlights include the incorporation of new tools and resources into the MGnify[5] and MetaPIPE[6] analytical pipelines (e.g. MAPseq, ITSOneDB),[7,8] the formal description of the MGnify pipeline using the common workflow language (CWL[9]) to promote interoperability, the establishment of the Marine Metagenomics Portal and MAR databases,[10] and a community paper (beyond ELIXIR) promoting best practices advocating the use of community standards for contextual provenance and metadata at all stages of the research data life cycle.[11] Capacity building has also been an important activity since the establishment of the Community, and many hands-on workshops and training courses have been developed and completed to build competence and expertise in a broader marine academic community.

However, the popularity of metagenomics has continued to grow, with current approaches providing greater genome-resolved insights into the community composition and the functions performed by the microbial constituents, with annotations spanning viruses, bacteria, archaea and microbial eukaryotes.[12–16] Furthermore, metagenomic-like approaches are increasingly being applied to untangle complex holobiont genomes such as lichens, where both the

primary symbionts and secondary non-obligate microbes are captured.[17] Finally, multi 'omics datasets are now being more routinely produced to understand not only the genetic potential, but also the actively produced transcripts, proteins and/or metabolites, with a view to establishing the links between genotype and phenotype. When a host organism is involved, such datasets can also be augmented with genetic data from the hosts, such as genome, single nucleotide polymorphisms and transcriptomic data. The collective data facilitate a hologenomic approach[18] to understanding host phenotypes, in the context of their environment and microbiome. Given this increasing complexity of study designs, and the broad applicability of microbiome research, we advocate expanding the Marine Metagenomics Community to include other areas of microbiome research. In particular, we highlight the need for an ELIXIR Microbiome Community to develop and promote standards and research infrastructures that enable the sharing of efforts, concepts, and best practices, while benefiting from the synergistic interplay with other ELIXIR Communities.

The scope of the ELIXIR Microbiome Community

The term metagenomics is often colloquially applied to many different areas of microbiome research (see Table 1), regularly (incorrectly) used to encompass both shotgun metagenomics (indiscriminate sequencing of DNA from an environmental sample) and metabarcoding approaches (the sequencing of a specific amplified marker gene). Depending on the nature of the scientific question being addressed and/or the environment, metagenomic analysis may also involve assembly, and potentially the generation of metagenome assembled genomes (MAGs).[19] Equally applicable is the analysis of unassembled raw-read data sets that can be used for taxonomic classification (e.g. Kraken,[20] MetaPhlan,[21] mOTUs[22]) and functional profiling approaches that are especially effective when extensive reference databases are available. Emerging sequencing technologies such as long-read sequencing methodologies and the associated adaptive sequencing techniques,[23] together with changing protocols such as host material depletion protocols (e.g. Ref. 24), are facilitating the analysis of a wide-range of differing communities. However, the applicability of certain downstream processing and/or analysis tools changes fundamentally in these different contexts. Similarly with meta-genomic data, metatranscriptomic data can be processed in different ways, and with an associated metagenomic dataset from the same sample, enables the estimation of both the genetic potential and actively transcribed fraction. Additionally, metaproteomics, an emerging technology in microbiome research, involves the study of the entire complement of proteins expressed by microbial communities in a given environment. Fundamentally, the ELIXIR Microbiome Community is about providing the necessary infrastructures required to perform analysis of nucleotide sequence data derived from a microbiome, especially the reproducibility of the results, the archiving and discovery of analyses and the interoperability of tools and data. Given the heterogeneity of such nucleotide data, this Community will work with other ELIXIR communities to determine how microbiome-derived data coming from different 'omics approaches, may be processed and integrated.

A fundamental challenge for this Community to address will be the provision of infrastructures that are sufficiently adaptable to permit the most appropriate informatics analysis, depending on the environment sampled and the experiments conducted. Moreover, when wishing to contextualise the results with similar experiments, the way a dataset has been produced and processed must be transparent to establish whether it is comparable (e.g. amplified sequence variants can only be compared when the same amplified regions are compared). Furthermore, when different methods are applied, best practices in data stewardship are required to ensure that the connectivity of the derived sequence data products, together with functional and taxonomic assertions are kept in context of the original sample/sequencing effort and associated contextual metadata. Finally, and possible unique to this ELIXIR Community, is the variety of researchers wishing to undertake microbiome research, spanning clinicians aiming to understand the role of the human microbiome in disease aetiology, ecologists wanting to understand the changing landscape of biodiversity, the agritechnology sector wishing to enhance animal and crop production, to biotechnology scientists looking for novel enzymes, among others.

Table 1. Overview of the terms and techniques used to study microbiome samples.

Term	Definition
Metabarcoding	Amplification and sequencing of diagnostic marker gene(s) found in a microbial community
Metagenomics	Random sequencing of the total DNA found in a microbial community
Metatranscriptomics	As with metagenomics, but the sequencing of the total RNA
Metabolomics (non-targeted)	Indiscriminate study of small molecules and products of metabolism
Metaproteomics	Identification and quantification of proteins found in a microbial community
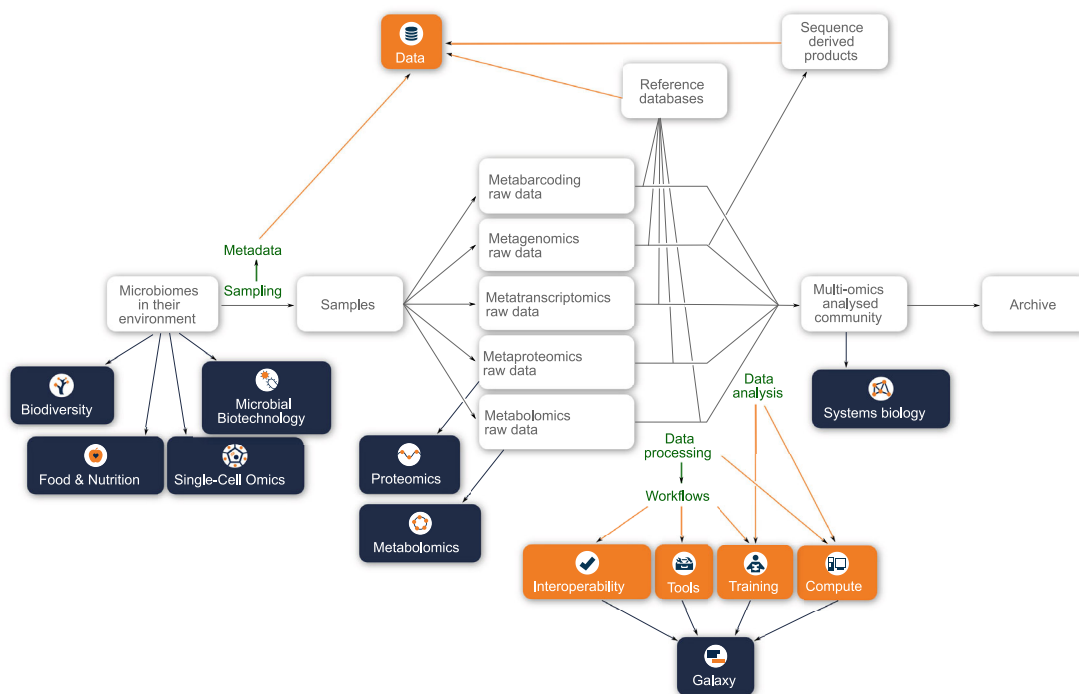
Figure 1. A schematic of how a microbiome sample (i.e. community in the environment) may be analysed using different 'omics approaches, with the main steps indicated in green. Underpinning these analyses will be the metagenomic and metatranscriptomic data, which will be used as a framework for the metaproteomic and metabolomic interpretation. Highlights in this figure are connections with the ELIXIR platforms (orange boxes) and other ELIXIR communities (dark blue boxes).

The context within ELIXIR

Given the breadth of aforementioned applications of microbiome research, it is unsurprising that there are many links to other current and future ELIXIR activities. Figure 1 presents a schematic layout of the experimental design of a multi 'omic analysis of a microbiome sample. Even in this very high-level representation, it can be easily observed that the new ELIXIR Microbiome Community has many potential interactions with other ELIXIR Communities and Platforms along the experimental workflow. Thus, the microbiome Community represents a showcase of the essence of ELIXIR by bringing together diverse informatics infrastructures that can be coupled together (interoperate) to achieve complex data analyses (on compute infrastructures) that have the appropriate provenance, with data adequately archived in the relevant ELIXIR core data resources. At all levels in ELIXIR, it will be essential to coordinate activities to ensure functional harmony between ELIXIR Communities using Platform-devised solutions.

Interactions with other ELIXIR Communities

Some of the key areas of interactions, both ongoing and foreseen, with other Communities are listed in Table 2. As indicated in Figure 1, the interaction with other ELIXIR Communities, specifically those concerned with environmental sampling, begins at the start of the data lifecycle, concerning the sample acquisition and characterisation of the microbial communities. For example, the Food and Nutrition Community aims to understand the relationship between food choices and human health. While microbiome analysis forms part of this Community's activities, the aim of the Food and Nutrition Community is to integrate microbiome data within the context of food and nutrition data, host genotype and phenotype information, and develop interventions that may impact disease.[24,29] Thus, in the case of the Food and Nutrition Community the microbiome is only a small part of the overall research program, and restricted to human microbiome research. Members of the existing ELIXIR Marine Metagenomics Community are already engaged with the Food & Nutrition Community, and have helped to provide microbiome sequence analysis services. Similarly the Biodiversity Community has multiple overlapping activities, but with a distinct remit. For example, computational infrastructures and tools borne out of metagenomics research are now being applied for pathogen and biodiversity surveillance. Furthermore, taxonomic inventories resulting from analysis of metagenomic/metabarcoding data are commonly accepted as biodiversity resources and biodiversity resources such as GBIF and OBIS[30] routinely incorporate data from both MGnify and International Nucleotide Sequence Database Collaboration (INSDC). Similarly, many of the biodiversity approaches use marker gene amplification for studying environmental DNA (eDNA). While this can overlap

Table 2. Overview of existing and planned interactions with ELIXIR Communities.

Community	Existing and planned interaction
3D BioInfo	Increasing numbers of structural models from metagenome-derived proteins are being produced (e.g. ESMAtlas).[25] Improving the organisation, quality control and presentation of these models will be essential to understanding structure-function relationships and improving the functional annotations of microbiomes.
Biodiversity	Connecting biodiversity/observation resources with 'omics data/analysis. The metabarcoding pipelines have many similarities with those used for eDNA analysis. Accessing the genomes of novel biodiversity presents similar annotation issues as faced by the Microbiome Community.
Federated human data	The field hotly debates whether human microbiomes should be considered as personal and sensitive medical data. Currently, the opinion and legislation varies from country to country. The use of solutions from the Federated human data Community for sharing sensitive data may become essential to the Microbiome Community.
Food & Nutrition	How metagenomics techniques can be used to understand the role of the gut microbiome in unlocking nutrients in food. We will provide standard analysis workflows to this Community.
Galaxy	The Galaxy Community operates across Platforms and Communities. Galaxy and its community provides a graphical interface to tools, a strong workflow manager to build workflows, computational resources with the European Galaxy server, and a powerful infrastructure and resources for training. Together with the Galaxy Community, we will continue to work to tailor and expand these resources to the Microbiome Community, given the needs identified by the ongoing evaluation study.
Metabolomics	Develop methods and tools to connect metabolic capabilities (metagenomics) and activities (metatranscriptomics and metaproteomics) of organisms/communities with effective presence/absence of the resulting products (metabolites).
Microbial Biotechnology	Improving the identification of valuable enzymatic activities from environmental genomics data to identify bioactives (e.g. enzyme, small molecule) of interest for the bioeconomy, for applications in food preservation, agriculture, chemistry or medicine.
Plant Science	Microbes play a key role in plant health and disease. Developing a greater understanding of the needs of the Plant Science Community for microbiome-based solutions to improve plant resilience and combating pests, as well as understanding how plants maintain their microbial communities across generations.
Proteomics	The interpretation of (meta-)proteomics spectral data requires sample-specific reference databases. Enable the production of tailored reference databases (e.g. biome and/or other contextual metadata) and the integration of metagenomic, metatranscriptomics and metaproteomics results.
Single Cell Omics	Single amplified genomes (SAGs) present an alternative strategy for understanding microbial communities. There are key areas of overlap in data standards,[26] with common issues on taxonomy, gene calling and protein functional annotation. Spatial single cell data is also improving the quality of MAGs/SAGs[27] and enabling the identification of microbes in tissues and tumours.[28]
Systems Biology	Empowering a better integration of multi-omics environmental approaches at the community- and environment-level to describe and understand how different community members interoperate to achieve processes.

with the metabarcoding approaches used in the Microbiome Community, eDNA analysis extends to marker genes such as Cytochrome c oxidase subunit I (Cox1) that is specific to macro-organisms and, thus, out of scope for the Microbiome Community and falls into the realm of the ELIXIR Biodiversity Community. Isolation of genomes, sequencing and their annotation is another area under the Biodiversity Community, but will encompass both microbes and macro-organisms. Nevertheless there will be common approaches used by this Community and the Microbiome Community for de novo annotation of novel genomes.

With the growing number of multi 'omics datasets, establishing strong ties with the ELIXIR Metabolomics and Proteomics communities[31] will be essential for understanding how metagenomic and metatranscriptomic data may be utilised by these Communities (e.g. the production of reference databases), and the nature of the data types produced by these other 'omics technologies, their limitations and how the data could be integrated. For example, overlaying metabolomics results on metagenomic data is currently non-trivial due to the scarcity of small molecule annotations that can be linked to functional annotations. Ongoing work with the Microbial Biotechnology and Systems Biology

Communities has identified the need to augment the functional annotation of metagenomic and metatranscriptomic data with chemical reaction information from resources such as Rhea.[32] While this will improve the discovery of new industrial applications, there is still the need to expand the protein functional annotations of the, so-called, microbial dark matter. The advent of new structural modelling software[25,33] and data resources[25,34] means that there are now structural models for millions of proteins that currently lack functional annotations, yet appear structurally related to functionally characterised proteins. Connections to the 3D BioInfo Community will aid how we store and organise this structural model information, reuse software components for visualisation and leverage their training materials on how to interpret structural model data. This will allow the Microbiome Community to assess the merits and limitations of this data type.

In summary, there are many synergies and connections between the Microbiome Community and the other existing ELIXIR Communities, but none of these Communities are focused on the core issues concerning microbiome-derived sequence analysis, infrastructure provision, data standards and best practices. Moreover, there are key societal challenges such as food security, climate changes, antimicrobial resistance (and new therapeutics) and pandemic preparedness where microbiome research plays a role, yet each one of these areas is far greater in scientific scope and therefore requires the collective outputs from more than one ELIXIR Community, and reach far beyond informatics research (Table 3).

Table 3. Description of current national and pan-European efforts aimed at microbiome research of relevance to the ELIXIR Microbiome Community.

Initiative	Acronym	Country	Aim
Mutualised Digital Spaces For Life Sciences	MuDIS4LS	FR	The main goal of MuDiS4LS is to develop a framework that will rely on the national and regional data centres to enable scientists controlling the flow of biological data, from their origin (data-producing national infrastructures) to their public release in national or international repositories, while ensuring their mid-term security during the intermediate phases of analysis and exploitation.
microGalaxy		Global	microGalaxy is a community of practice to (i) develop and sustain microbial data analysis in Galaxy, (ii) implement standardised "best practices", (iii) expand documentation and training, (iv) coordinate efforts in tools, workflows and training development
NFDI4Microbiota		DE	The mission of this consortium, part of the German NFDI (National research Data Infrastructure), is to be the central hub in Germany for supporting the microbiology community with access to data, analysis services, data/metadata standards and training. The main aims and objectives of NFDI4Microbiota are: (i) generate a broad awareness of the importance of the FAIR principles, open science and reproducible research in the microbiological community and drive a cultural change toward their widespread adaptation; (ii) equip the community with the required skills and literacy for efficient and data-driven microbial research by providing a comprehensive training program; (iii) improve the research process by mobilising, structuring and linking available data, information and knowledge related to microorganisms; (iv) support high-quality research data management by introducing professional data stewards into the microbiological research process; (v) increase the value of data by standardising and systematically collecting rich metadata and building tools for querying; (vi) make research more reproducible by standardising data processing and analysis; (vii) provide computational tools and infrastructure for the translation of data into new knowledge.

Table 3. *Continued*

Initiative	Acronym	Country	Aim
European Reference Genomes Atlas	ERGA	Europe	The European Reference Genome Atlas (ERGA) initiative is a pan-European scientific response to current threats to biodiversity. Reference genomes provide the most complete insight into the genetic basis that forms each species and represent a powerful resource in understanding how biodiversity functions. With approximately one fifth of the ~200,000 European species at risk of extinction, we need to act fast and together to generate high-quality complete genome resources on a large scale.
Metaproteomics Initiative[35]			Promoting dissemination of metaproteomics fundamentals, advancements, and applications through collaborative networking in microbiome research. They aim to be the central information hub and open meeting place where newcomers and experts interact to communicate, standardise and accelerate experimental and bioinformatic methodologies in this field.
Secured computing spaces for the data access and analysis project of the France 2030 programme « Food Systems, Microbiome and Health »	Cloud4SAMS	FR	The Cloud4SAMS targeted project aims to deploy a distributed digital infrastructure enabling researchers to exploit microbiome and health data in a secure computing environment. It relies on the federation of computing resources operated by different institutions and spread over different sites: datasets produced by microbiome projects (in particular those to be funded by the France 2030 programme), software tools and workflows for processing these data, computing and storage platforms suitable for processing microbiome data and matching them with health data. These resources will be indexed in the Cloud4SAMS catalogue, and will serve as building blocks to define deployment recipes describing all the procedures to instantiate a virtual machine in a secure cloud, to install the whole software environment and to transfer the datasets needed for the project. Access to these data is facilitated by an interface that manages the requests to the access committees of each project, the validation of authorizations, the ad hoc extraction of the data and their transfer to the secure spaces.
Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics	ELIXIRxNextGenIT	IT	ELIXIRxNextGenIT is a national project, in continuity with CNRBiOmics, aimed at consolidating the ELIXIR-IT Infrastructure for Omics and Bioinformatics. The project is focused on data production, computational analysis, facilities improvement and human resources recruitment and training, with a view to increasing the national ELIXIR Infrastructure potential, including the capability to host new resources such as the Federated European Genome-phenome Archive (FEGA).
European e-Science Infrastructure for biodiversity and ecosystem research	LifeWatch ERIC	EU/IT	The project aims to accelerate the sharing, integration and analysis of open-data and its Virtual Research Environments (VREs) to enable studies on biodiversity structure and conservation related to multiple drivers. The LifeWatch Italy Joint Research Unit (JRU) coordinates the Italian contribution to LifeWatch ERIC, the national activities of the LifeWatch Service Centre and the LifeWatch-ITA distributed e-Biodiversity Research Institute that includes the Biomolecular, Collections, Interactions and Mediterranean thematic centers.

Table 3. *Continued*

Initiative	Acronym	Country	Aim
National Research Center in Bioinformatics for Omics Sciences	CNRBiOmics	IT	The project aims to enhance the ELIXIR Italian node infrastructure mainly in the southern regions. With its headquarters in Bari (Apulia region), it is engaged in the establishment of a "centre of excellence" for 'omics data production, management, and analysis. The most advanced laboratory platforms for second and third generation sequencing, proteomics, metabolomics and transcriptomics are integrated with computing and storage high power platforms situated in the Bari hub and interconnected with the existing ELIXIR infrastructure. The establishment of an higher education training platform to provide the necessary skills for the infrastructure optimal use is also envisaged.

Similarly, microbiome research has many translation aspects, ranging from the discovery of biomarkers associated with health and disease to industrial applications such as using enzymes from microbes or the microbes themselves for performing bioremediation and/or replacing chemical processes. One topic that is an area of intensive research is the discovery of enzymes capable of degrading plastics, typically polyethylene terephthalate (PET).[36] While metagenomic assembly and analysis is providing a rich source of new enzymes, the informatics at the core of the Microbiome Community will not provide the information why one enzyme should be assayed in preference to another, how these alpha-beta hydrolases have adapted to utilising PET, or why one enzyme performs better than another. Such answers will come from the collaborative efforts that bridge across Communities, such as microbial biotechnology and 3D BioInfo and, of course, the wider research community.

Interaction with ELIXIR Platforms

Similar to the collaborations with the ELIXIR Communities, there are multiple ongoing and future interactions with the ELIXIR Platforms. In the following sections the connections between the past Marine Metagenomics Community or the future Microbiome Community and each of the Platforms will be highlighted.

Data

The aim of the ELIXIR Data Platform is to promote the use, re-use and value of life science data. A key part of this activity has been the establishment of the Core Data Resources (CDR). Underpinning sequenced-based microbiome research is the European Nucleotide Archive (ENA), which is a recognised CDR and part of the INSDC, which in collaboration with the National Institute of Genetics DNA DataBank of Japan (DDBJ) and the United States National Center for Biotechnology's (NCBI) GenBank and Sequence Read Archive (SRA), facilitate the deposition and global exchange of sequence data. Alongside the archived sequence data, users can access comprehensive metadata that is important to contextualise where the data originated. Throughout the lifetime of the ELIXIR Marine Metagenomics Community there have been extensive efforts to increase the standardisation of derived sequence products from metagenomic short-read datasets, particularly increasing the availability of assemblies[5] and the introduction of the deposition layers to support the increase in the numbers of MAGs being generated.[37] In the new Microbiome Community we will continue to promote and develop these layers to accommodate Eukaryotic MAGs (see below), viral sequences and complex coassembly, as well as incorporating the latest community standards as they are approved by authoritative bodies. The work undertaken to generate the MAR databases highlighted that many marine samples in ENA lack key metadata fields. Through extensive curation efforts, using literature as well as contacting the original data submitters, much of this missing data was retrieved and added to the MAR database. While ENA (or any of the INSDC partners) can not add this metadata to the original sequence record, an ELIXIR sponsored initiative led to the establishment of the Contextual Data Clearinghouse (CDCH). The CDCH facilitates the capture of additional metadata using controlled vocabularies including a description of how this data was generated (e.g. manual assertion, computationally derived), so that they can be associated with an INSDC record. Longer term, this data will be incorporated into BioSamples.

In other non-sequenced based 'omics fields, microbiome data archiving and analysis is supported by data-type specific resources. In the case of metaproteomics, the PRIDE database repository (also an ELIXIR CDR) enables archiving and re-analysis of (meta) proteomics data, and now also encourages researchers to upload their metadata in SDRF-format.[38,39] PRIDE is the leading resource of the International ProteomeXchange Consortium of proteomics data resources, involving additional databases in the USA, Japan and China, in addition to PRIDE. Similarly, in the case of metabolites the data can

be deposited in the MetaboLights repository[40] or similar resources. A current challenge facing the field is connecting different multi 'omics data that have been derived from the same sample.

The Data Platform also promotes the linkage between Europe PMC[41] and other CDR databases. This is critical for the Microbiome Community as additional contextual metadata can often be found in the literature,[42,43] providing crucial overarching context to the experiment, which can be important for meta-analyses. We will continue to promote such approaches, enriching metadata wherever possible.

Last but not least, new activities will be promoted aimed at the integration of microbiome data coming from different 'omics approaches. In this context, recently, the PRIDE and MGnify teams developed and implemented new pipelines in both platforms for the re-analysis and integration of metagenomic and metaproteomic data, allowing the re-analysis of metaproteomics datasets from PRIDE using sequence databases generated from MGnify, and contextualising the results back into the MGnify web interface in terms of assembly annotations (https://github.com/PRIDE-reanalysis/MetaPUF). The ELIXIR Microbiome Community will also work to move the Marine Metagenomics domain in the RDMKit towards a more general Microbiome domain.

Tools

Microbiome data analysis employs a large number of tools which are used to perform basic quality control on the sequence data, with separate tools (and reference databases) typically used for taxonomic and functional profiling. Installing and managing dependencies has been eased by the use of package management systems such as Conda, or through the use of containers, e.g. Singularity. The ELIXIR Microbiome Community will increase their use of BioContainers[44] to promote the packaging, containerisation and deployment of tools relevant to microbiome research.

In order to make tools findable the Community will work on improving their annotation by (i) expanding the EDAM ontology[45] to include microbiome-specific keywords, (ii) performing periodic reviews of tools and their associated annotations in the bio.tools[46] catalogue. These annotations will subsequently be used to build a catalogue of tools for microbiome data analysis and their availability for different platforms, e.g. Galaxy, or as workflow descriptions (e.g. Snakemake, CWL, Nextflow), which can be readily combined to make new annotation workflows. Additionally, the Community will develop and maintain cloud-deployable and FAIR analysis pipelines using state of the art tools and following best open science practices by: (i) using workflow descriptions; (ii) documenting the workflows and depositing them in WorkflowHub[47] for easy discovery, re-use and assessment; (iii) making them available for the Community via platforms such as MGnify and Galaxy.

As an integral part of the Tools platform, Galaxy has integration with OpenEBench, WorkflowHub EDAM, bio.tools and follows all Software Best Practices. A joint effort between the Microbiome and Galaxy Communities is running an evaluation of tool requirements for microbiome data analysis in the Galaxy ecosystem. This evaluation will lead to a shared roadmap between both Communities for tool integration and standardised workflow development for microbiome data analysis.

Benchmarking

Very few analyses in microbiome research employ a single tool, with the norm being the coupling of multiple tools and reference databases to achieve a comprehensive analysis that includes both taxonomic and functional results. Even relatively simple workflows that perform metagenomics assembly are computationally heavy. This combination of workflow complexity and typical computational overheads has always made the routine benchmarking tools for microbiome informatics research burdensome. Nevertheless, where two or more tools perform equivalent tasks, it can be relatively simple to modify existing formally described workflows to evaluate their respective performances, but that ease often depends on where they occur in the overall workflow and the metrics used to evaluate the tool. Many efforts have tried to compare the outputs of tools and workflows (e.g. Refs. 48–52), with the Critical Assessment of Microbiome Interpretations (CAMI) having become an internationally recognised benchmarking effort.[53–56] The CAMI challenges have established a range of benchmarking datasets for evaluating different categories of tools. Importantly, the organisers of CAMI have engaged data generators to provide data, such that truly independent benchmarking can be undertaken. However, these benchmark datasets can become outdated over time, as the underlying data enters the reference database. The Galaxy Community has already investigated implementing benchmarking infrastructure using CAMI datasets, and increasing the awareness of this infrastructure will be a key effort across Communities and Platforms. As the Microbiome Community establishes, we will develop a broader understanding of the requirements of the Community, feed this to the Tools Platform, as well as seek opportunities to interact with the Tools Platform to capture the diversity of tools and their utility via such benchmarking activities.

Compute

Given that most academic institutions have access to dedicated sequencing facilities or equivalent commercial facilities, coupled with the diminishing costs of DNA sequencing and other 'omics technologies, it is relatively easy to generate large datasets, but significant computational resources are required to store and analyse the data. Depending on the analysis being performed, the computational requirements can be very different. For example, metagenomic assembly typically requires small numbers of cores on a large memory machine, whereas some forms of raw-read analysis require many cores (hundreds) with a small memory footprint. As such, microbiome researchers need to understand the likely computational costs, and their options for deploying them on high performance computing (HPC) and cloud environments. Efforts such as Blue Cloud have helped reduce some of the barriers to using the European Open Science Cloud (EOSC) for marine research through the delivery of a collaborative virtual environment, but the range of services is limited. While such efforts help, there are still many barriers to accessing compute resources and deploying complex metagenomic pipelines in a distributed or even hybrid fashion. Working with the Compute Platform, the ELIXIR Microbiome Community will continue to investigate solutions that facilitate the execution of workflows within such distributed and/or hybrid environments, e.g. using Pulsar network, the distributed compute network offered by the Galaxy Community, and provide guidance of the likely costs of using compute infrastructures.

Interoperability

Previous work by the Marine Metagenomics Community has leveraged many of the ELIXIR Interoperability Platform solutions, especially the use of workflow languages for the formal description of pipelines, improving the provenance of the data outputs. As such, both the MetaPIPE and MGnify pipelines have been described using the Common Workflow Language (CWL). This effort was paralleled by MG-RAST, which also allowed MGnify and MG-RAST to exchange pipelines[57] and establish that the biological signatures reported by the respective pipelines were very similar, yet confounded by different reference databases and methodologies for assigning function. Since then, MGnify has published their workflows in WorkflowHub,[58] further promoting their discovery and reuse. As an example of reuse, the MGnify pipeline has been used as the basis for the newly developed metaGOflow pipeline,[59] to be used by the Marine Genomic Observatories. Moreover, this work also employed Research Object Crate (RO-crate)[60] to package relevant metadata about the sample and the bioinformatics analysis applied and the data products. RO-crate offers new opportunities for sharing or federating the metagenomics analysis workload. In parallel, Galaxy, which supports the Tool Registry Service (TRS) protocol to exchange and run workflows between the WorkflowHub and Galaxy, gained support for RO-Crate (version 23.0) to export complete data analysis as a structured and FAIR digital object, supporting the GA4GH standards, and is in the process of applying to be a Recommended Interoperability Resource.

The Microbiome Community will continue to work with the Interoperability Platform to make wider use of RO-crate, with a view to federate data analysis between resources. For example, future work by the new Community will enable the MGnify workflows to be made deployable on Galaxy, with the RO-crate to be transferred, verified and ingested into MGnify. Additional work needs to be undertaken to understand how universal this approach is, so that MGnify could become a hub for a range of additional analyses, thereby reducing the duplication of effort that currently exists in the community.

Finally, we will work on the development of novel mechanisms to integrate and link data coming from multi-omic approaches using different tools and data resources. This will require the development of new data Interoperability layers for data resources that are not normally focused in Microbiome data, such as the PRIDE database in the case of metaproteomics data.

Training

One of the key areas commonly highlighted by national and international reports on the potential of microbiome research is the need for training, especially in the area of informatics. As already highlighted, microbiome analysis is an emerging and evolving research field by itself, with plenty of challenges still to be addressed. Combined with this complexity, the increasing number of researchers using such methods makes the need for continuous training and re-training a challenge on its own. Researchers need to become familiar with modern computing technologies, such as HPC and cloud computing, and follow the constant updates on experimental approaches, algorithms (new and updates) and pipeline developments. As new pipelines are established and existing pipelines improved through the incorporation of new tools and/or reference databases, this adds further complexity to the tool and data output landscape associated with microbiome research.

Platforms such as MGnify support large-scale services for most, if not all, steps of a microbiome study, meaning the distribution of raw-data, production of assemblies, their analysis, and their potential use for meta-analysis, have proved of great benefit. Nevertheless, these analyses should be considered just the starting point for further downstream analysis,

which requires the specific domain expertise of the researchers involved in undertaking the study. One approach can be the use of cloud-based initiatives such as Galaxy supporting graphical interfaces and allowing the users to choose more specific tools, while tuning their parameters and reference databases according to their environment being studied. Such infrastructures attempt to fill the gap between researchers without experience in computer science and their needs for FAIR and quality microbiome analysis. Despite both solutions being readily available, there remains knowledge gaps and/or reticence about using such resources, often due to a lack of training.

To upskill microbiome scientists and keep them up-to-date in microbiome data analysis and standards, the ELIXIR Microbiome Community will work in coordination with the ELIXIR Training Platform to offer scalable and FAIR training. The Microbiome Community will continue to: (i) annotate training materials with appropriate metadata to create a comprehensive training portfolio; (ii) FAIRify the training content, making it open-access; (iii) register training material, national and international providers and events in ELIXIR's Training Portal TeSS;[61] (iv) assist the Training platform in the development of annual training gap surveys; and (v) develop materials and design learning paths specific to different community needs (e.g. biomes or data types).

To enable access to training resources and deliver this training, face-to-face and online workshops will be organised and videos will be recorded for "on demand" learning. The technical infrastructure for training, in particular the computational environment setup and software installation challenge will be addressed in coordination with the ELIXIR Tools and Compute Platforms, with the aim of promoting the use of Conda environments, containers, notebooks or platforms like Galaxy which mitigate many of the current obstacles. In order to make these aspirations possible, the Community will increase its training capacity by working with training communities on practices, organising Train the Trainers events and building a community of microbiome research trainers, with areas of expertise covering different environments, 'omics approaches and data analysis pathways. Ensuring these trainers maintain their knowledge with the evolving informatics landscape is, arguably, a key challenge that is yet to be addressed and something this Community will strive to solve in collaboration with the ELIXIR Training Platform.

Context with other international initiatives

We have highlighted the need for promoting best practices and standards throughout this article. However, it is also important that the Microbiome Community continues to build upon engagement with organisations such as the Genome Standards Consortium (GSC[62]). The GSC is an international organisation aimed at making genomic data discoverable through the establishment of standards, which are derived from community input. This consortium includes stakeholders from across the data life cycle, from research scientists producing the data, to data analysts to database providers. By engaging this range of stakeholders, the GSC have become critical for establishing many of the standards that underpin genomic research. Examples of GSC established standards that are particularly pertinent to the microbiome domain include: minimal information about any sequence (MIxS[63]), the Biological Observation Matrix (BIOM) format[64]; and the Minimum Information about a Metagenome-Assembled Genome (MIMAG[26,63]).

The GSC has many ongoing projects relevant to the ELIXIR Microbiome Community, especially the M5 project (Metagenomics, Metadata, MetaAnalysis, Models and MetaInfrastructure). Combining the activities on standards concerning workflows is critical for the global microbiome community to operate with a consistent and unified view on how to make microbiome analysis reproducible and stand-up to scientific interrogation. Note, such standards do not restrict what analysis can/should be performed, but rather provide the appropriate information, that given the same starting input data, exactly the same analysis, and hence result, can be achieved.

There are also other ELIXIR Node-specific initiatives that the Microbiome Community connects with to ensure that the respective efforts are synergised. Examples of projects with ELIXIR Node involvement directly related to the ELIXIR Microbiome Community are presented in Table 3, which cover a diverse range of topics. The engagement needs to be bi-directional to ensure that the needs of Nodes are well understood and that solutions developed at national levels can be spread across the ELIXIR Microbiome Community, and vice versa. In this context, the ELIXIR Microbiome Community leads will undertake coordinating roles, engaging with the project representatives, inviting them to relevant ELIXIR events and promoting active participation in relevant ELIXIR Communities.

MicrobiomeSupport, formerly a European Commission funded Community Action Support program aimed at improving microbiome research and innovation, highlighted in their final report[65] that there was "limited connectedness" in microbiome research conducted on different environments/systems, and that during the course of this program the lack of connectedness did not improve. This independent finding reinforces the need for broadening the ELIXIR Marine Metagenomics Community to a more generalist Microbiome Community.

It will also be important to showcase the ELIXIR Microbiome Community to European countries that are yet to join ELIXIR. For example, Romania has a thriving microbiome research community, but is faced with the same set of informatics challenges. Sharing knowledge beyond ELIXIR, will not be the primary goal, but will nevertheless be important to harmonise the activities internationally and promote the benefits of participation in ELIXIR. Beyond Europe, there are parallel organisations that strive to achieve similar goals to ELIXIR in other locations. For example, Australia BioCommons aims to promote bioinformatics and bioscience data infrastructures at a national level. Given the strength of microbiome research in Australia (see below), we will explore opportunities for international collaboration.

In addition, it will be important to showcase the ELIXIR Microbiome Community to communities (within and outside Europe) that are not yet familiarised with ELIXIR activities. For example, the Metaproteomics Initiative is an international community that promotes dissemination of metaproteomics fundamentals, advancements, and applications through collaborative networking in microbiome research.[35,66] For example, recently, they benchmarked metaproteomics workflows and bioinformatics methods in the field in the first multi-lab benchmark study in metaproteomics (called CAMPI), showcasing the robustness of metaproteomics data analysis workflows.[66]

Finally, the National Microbiome Data Collaborative (NMDC),[67] a US led initiative, is developing a unified data portal to support microbiome multi-omics data integration and analysis through an integrated, distributed framework. Many of the governing principles associated with this portal are common with those described here, especially with the desire to have containerised, reusable computational workflows, as well as trying to make the data compliant with the FAIR principles. Sharing experiences and best practices between NMDC and the ELIXIR Microbiome Community (and others) will improve the global standardisation of microbiome research.

Interaction with other key data resources beyond ELIXIR

Microbiome research is global, so it is also key that European microbiome research infrastructures are coordinated with other international resources. Below we highlight a small selection of widely used resources that are produced outside Europe, and place them in context of the ELIXIR Microbiome Community. Some of the most utilised tools and resources used by the current Microbiome Community are CheckM,[68] the Genome Taxonomy Database (GTDB) and the associated GTDB toolkit.[69,70] CheckM is widely used to assess the completeness and contamination of prokaryotic MAGs, and is part of the GSC reporting standard. The GTDB resources is a genome based taxonomy of prokaryotes, and the associated GTDB-tk facilitates the classification of other prokaryotic genomes against this framework, more often than not, to determine novelty. These are currently made available via the Australian research groups, who face similar challenges in maintaining resources. Other key resources are based in the US, with MG-RAST[71] produced by Argonne National Laboratories and a range of different resources produced by the Joint Genome Institute (JGI). MG-RAST facilitates the analysis of raw-reads and assemblies (metabarcoding, metagenomics and metatranscriptomics), but does not perform assembly nor offer any form or long-term archiving assurance. The JGI IMG/M resource[72] has many parallels with MGnify, offering a wide range of data analyses focused on assembly and MAG generation, but IMG/M does not deal with metabarcoding. Notably, JGI also produces IMG/VR,[73] a globally unique collection of viruses, many of which have been determined from metagenomic and metatranscriptomics. Any future effort in Europe focused on viruses must aim to minimise the duplication of effort and content with IMG/VR. Coordinating with these global initiatives is key to ensure the future availability of the tools and resources, ensuring interoperability between the resources, maintaining uniform standards and sharing of the informatics/computational burden.

Specific challenges and objectives of the ELIXIR Microbiome Community

A key early challenge in developing the ELIXIR Microbiome Community is to establish a detailed understanding of the current approaches and databases used for the analysis of different microbiomes. For example, it is widely accepted that current short-read assembly-based methods do not generally work as well for soil microbiomes due to the diversity of the microbial community typically present (the sequence depths are insufficient to build useful contigs or the datasets are so large, that they are computationally intractable). This current limitation, has led and will continue to lead to the development of new experimental methods, from sampling through to nucleic acid sequencing and informatics analysis. In this section, some of the key challenges associated with microbiome research are highlighted below, together with how these challenges will be addressed by the new ELIXIR Microbiome Community. Table 4 lists the key thematic areas and objectives that the Microbiome Community will address, split into short-term and longer-term objectives to provide a high-level overview of the proposed Community activities.

The Microbiome Community will also provide a mechanism for sharing knowledge about new approaches for microbiome research, be it experimental or informatics-based techniques. For example, there is an increasing number of metagenomics datasets that are produced using long-read sequence technologies. While long-read sequencing technologies can require larger quantities of DNA or may be more error prone compared to third-generation short-read sequencing technologies – which can limit their use – the long-reads can mitigate the computational burden of

Table 4. Objectives of the ELIXIR Microbiome Community.

Area	Objective
Near-term (2 years timeframe)	
Community Expansion	Survey of needs, key datasets, data analysis approaches, 'omics data types and biome specific specialisation
	Identify key experts involved in viral, prokaryotic and eukaryotic analysis
	Establish and share a strategic technical roadmap with the Communities and Platforms, highlighting key contacts
	Identify relevant funding calls, with the aid of building microbiome research informatics capacities and connecting to key experts in other 'omics (e.g. metaproteomics)
Training	Increase awareness of microbiome tools, resources, and their applicability to different microbiomes
	Address knowledge gaps in generating and adopting workflows
	Advanced containerisation and cloud deployment
Co-ordinate	Increase rates of data archival deposition, with rich contextual metadata. Establish a mapping between biome and checklists
	Data analysis through the use of services
	Sharing of ideas on the design and implementation of workflows for microbiome research, promoting the use of best practices
	Organise in-person and virtual meetings for the Microbiome Community
Industry connection	Microbiome research has many applications suitable for pharmaceutical and biotechnological applications. Use ELIXIR and Node forums to understand demands and current limitations impacting this sector.
Longer-term (~3-5 years)	
Training	Targeted training for different microbiome communities
	Addressing the issue of maintaining "Train the Trainer"
	Organise hackathon to improve integration of ELIXIR services providing microbiome data
	Establish a rich set of training materials, appropriately tagged to aid discoverability
Federated data analysis	Enable the execution of MGnify pipelines in Galaxy and/or other ata management workflows, and submission of results to MGnify
	Establish routine mechanisms for federating microbiome analysis (e.g. RO-Crates, resources).
	Demonstrate approaches to multi'omics integration, through collaborative, cross-Community initiatives
Promoting new approaches	In conjunction with GSC, establish new standards for microbiome research, particularly with respect to data analysis reporting and contextual metadata reporting
	Leverage new data-types and experimental approaches to improve the scope and/or quality of microbiome analysis
	Enhance existing or establish new reference databases in response to Community demand and capacity
	Established new methods for across study comparisons, mitigating against confounding factors to enhance discovery
	Provide a mechanism for estimating the cost/benefit of performing different types of analysis in the context of different microbiomes
International harmonisation	Represent the Microbiome Community at international conferences, promoting Community/ELIXIR outputs and solutions
	Foster international collaborations between other resources providers and databases to ensure global harmonisation of e-infrastructures for microbiome research
	Leverage the CAMI initiative to facilitate benchmarking of tools and workflows

metagenomic assembly and increase the confidence in analysis results (e.g. MAGs produced by long-reads can have high contiguity and therefore less prone to contamination). The long-reads can be paired with short-read sequences, which can then be used in different ways (e.g. sequence error correction). Increasing the awareness of these long-read and hybrid-sequencing approaches, the workflows that support their analysis and when and where they could be applied will be a key output of the Microbiome Community. Similarly, there are other experimental approaches such as single amplified genomes (SAGs), which have increased in popularity. The Microbiome Community will also be important for assessing the utility of emerging sequencing approaches, such as adaptive sequencing approaches. In this case, the methods can access low abundance microbes, although such methods will not facilitate the generation of abundance profiles. Bringing these data types alongside the ubiquitous short-read datasets will require new standards and data integration approaches to be developed by the Microbiome Community.

There has been a paradigm-shift in metagenomic analysis with a common goal now being the generation of prokaryotic MAGs, which has not only allowed the identification of thousands of specific functions, but facilitated them to be assigned to specific organisms. As such, this has started the development of specific MAG deposition layers,[74] and the development of MAG specific resources. The new Microbiome Community will promote the use of MAG deposition, and provide guidelines and software to aid their deposition. Workflows that encompass both MAG generation and quality verification will be developed that include the capture of both prokaryotic and eukaryotic MAGs. The Microbiome Community will help establish best practices for eukaryotic MAG discovery, as well as develop new standards for removing redundancy and methods for assigning taxonomy, which are recognised gaps in the area of eukaryotic MAG discovery. While prokaryotic MAG recovery methods are more mature and standardised, it is anticipated that there will be continuous improvements in both experimental and computational methods for generating longer contigs, and more datasets that enable different approaches to enhance the detection of contamination and/or misassembly. The ELIXIR Microbiome Community will also evaluate methods and establish best practices for the identification of sub-species/ strains in metagenomic datasets. To do so, we will engage with efforts such as the Critical Assessment of Metagenome Interpretation (CAMI)[53,75] to identify tools that can scalably and accurately classify MAGs at a finer grain taxonomic level than species.

Finally, the classification and naming of MAGs is going to be paramount, so that the novel biodiversity can be understood and more easily referenced by the scientific community. Currently, the Microbiome Community has widely adopted the GTDB[69] and the associated GTDB-tk[70] for classifying MAGs against a reference tree. However, the taxonomy of GTDB differs from the more widely-used NCBI taxonomy, and there is a need to increase the interoperability between these two taxonomies. The ELIXIR Microbiome Community will work on addressing the current issues associated with MAGs and taxonomy. Additionally, another key area of development will be increasing the linkage between genomic resources and marker genes, such as the ribosomal small subunit (SSU) RNA.

In addition to cellular microbes, another area for the ELIXIR Microbiome Community to address is the development of the infrastructure and resources for identifying and cataloguing viruses in metagenomic and metatranscriptomic data.[76–79] Viral genomes are incredibly diverse in terms of composition and organisation. Viruses, particularly those that infect bacteria, are found ubiquitously in all environments and play critical roles in community dynamics. However, there are three challenges associated with viral microbiomes: (i) there is no universal marker gene covering all viruses; (ii) viral taxonomic frameworks are incomplete; (iii) there is no centralised database collecting the millions of viral sequences; and (iv) metagenomics informatics often only produces fragments of viruses, which causes ambiguities concerning their classification and functions. It will be critical for the ELIXIR Microbiome Community to engage with established viral infrastructures and organisations, such as the European Virus Bioinformatics Center, to establish methods, standards and resources for improving the analysis of viruses found in microbiome sequence data.

The increase in metagenomic assemblies has resulted in a parallel increase in the number of protein sequences that have been identified, with sets of non-redundant proteins now in the billions. There is huge potential for discovery in these protein datasets, as well as de novo designs fit for purpose, e.g. carbonic anhydrases[80] and a key aim for the new ELIXIR Microbiome Community will be ensuring that these data are annotated, both as individual sequences or as higher order grouping (e.g. pathways, biosynthetic gene clusters). This will involve the evaluation of emerging tools, as well as harnessing structural models to allow the detection of relationships that are undetectable by current sequenced based methods. The Community will need to work together to shed light on the functions of the so-called 'Dark Matter', develop standards for functional labelling that encapsulate both the mechanisms and confidence of the annotation, and develop new infrastructural frameworks for accessing slices of the data based on the requirements.

As identified by the ELIXIR Marine Metagenomics Community, experimental and contextual metadata is critical to comparative metagenomics. The absence of rich contextual and experimental metadata limits data reuse and the

production of downstream data products, such as assemblies and MAGs. With the expanded Microbiome Community, we will identify areas where metadata standards need to be improved, with biome specific contextual metadata being the most likely source of specific metadata checklist. The Community will develop training promoting the need for metadata, checking compliance against standards, how the metadata can be captured and submitted to accompany the sequence data, and potentially other 'omics data types. Within the Community, we will develop and promote standards around the analysis provenance (analytical metadata), and how the collective corpus of metadata can be used to improve meta-analysis and the identification of confounding factors when comparing different research projects.

Another key challenge that the Microbiome Community needs to address is ensuring that compute resources are accessible for performing the different forms of data analysis that can be associated with microbiome derived sequence data. Previously, we have highlighted the need for interaction with the ELIXIR Compute, Interoperability and Training platforms, as well as ELIXIR Communities such as the Galaxy Community. This requires that analysis pipelines are readily discoverable and deployable, and that key issues regarding both compute processing and storage requirements are well understood. Additionally, given that microbiome associated data analysis has such computational overheads, it is vital that models for data archiving and/or sharing are developed by the Microbiome Community to increase the capacity of microbiome research within Europe. This may require the development of new or extensions to existing databases, but it requires an agreement from the research community to adopt them. Achieving this will involve both communication and training of the microbiome research community.

While there are data resources such as MGnify that provide access to consistent analyses pertaining to different metabarcoding, metatranscriptomics and metagenomics datasets from a variety of biomes, it is fundamental to remember that these data outputs do not represent the end of the analysis pathway. Typically studies require comparison between different cohort groups (disease vs health, treatment vs non-treatment). Furthermore, as the biological signal from meta'omics datasets can be extremely noisy, there can often be the need to combine datasets to boost statistical significance of the biological signal. Similarly, the combination of studies can also be used to: (i) contextualise against previous studies (e.g. similar studies on the same diseases); (ii) understand the distribution of microbes or functional features (e.g. antimicrobial resistance genes) between different geographical locations; and/or (iii) study the relationship between biomes (e.g. studies adopting a OneHealth approach). To enable such large, complex studies there needs to be a greater understanding of the approaches suitable for cross study comparisons, and their limitations. Thus, a major objective for the Microbiome Community will be to include those researchers that are developing methods that can identify and mitigate experimental and informatic confounding factors, which currently limit data reuse. Existing approaches often rely on correlating contextual and experimental metadata with statistically significant factors identified in the datasets. There is also the need to develop and promote methods for performing robust statistical analysis of microbiome derived data, thereby enabling biological signals to be extracted from cross-sample/project datasets. Currently, there is a tendency to analyse the different 'omics datasets independently, and then correlate the derived signals. However, statistical methods are being developed to facilitate the analysis of integrated multi-omics datasets, and it will be important that the Microbiome Community determines the applicability of these approaches for microbiome research.

In the context of other 'omics approaches, there are also some major challenges in metaproteomics.[81] One of the major challenges is the construction of tailored protein sequence databases which are needed to identify proteins in complex microbial communities. Metaproteomics aims to elucidate the functional and taxonomic interplay of proteins in microbiomes, but the diversity and vast number of unknown and uncharacterized proteins present in these communities makes database creation and accurate protein identification difficult. As microbial communities are highly dynamic and their protein expression can vary significantly, conventional protein sequence databases might not cover the entire diversity, leading to potential limitations in accurate protein identification (e.g. the use of de novo sequencing). Addressing this challenge is crucial for improving the reliability and confidence of metaproteomic analysis and obtaining comprehensive insights into the functional roles of proteins in complex microbiomes.

As metagenomic methods have become a more routine method for studying microbial communities, metagenomics has been and will continue to be paired with more and more diverse sets of measurements of the microbiome. Examples of non-omics data collected alongside metagenomics data include geochemical (e.g. PANGAEA[82]) measurements, meteorological, image data and even acoustics. While methods are already emerging for the integration of 'omics datatypes (e.g. MOFA,[81,83] MIA), integration of these additional non-omics data types will enable a broader under-standing of microbiomes in context. For the new Microbiome Community, it will be essential to identify the appropriate archives for these data types, and establish the methods to facilitate navigating between datasets from the same samples. Only through achieving this, can new data visualisation schemas that enable the combination of environmental, geospatial and temporal data, in addition to biological data (taxonomy/function), be developed.

Conclusions

The overarching aim of the Microbiome Community is to develop a sustainable bioinformatics infrastructure for microbiome resources (data, tools, workflows, standards, training) which will enable a deep understanding of the function and taxonomy of the entire microbial fraction. We aim to be biome-agnostic, yet balanced in supporting the analysis and interpretation of data from different environments. We aim to highlight the very best approaches for the analysis and integration of different data types (e.g. sequences, metabolites, proteomics, and images) and their visualisation. By broadening the Community we will engage many more researchers and aspire to have a greater representation of scientists from different disciplines, such as ecologists and clinicians, complementing the strong molecular biology and genomics backgrounds already represented in the Community. The Microbiome Community will have key roles in engaging with policy makers (e.g. access and benefit sharing, climate change impact assessment), as well as the industrial sector, which is increasing the translation of basic research to microbiome-based products (e.g. UK Microbiome Strategic Roadmap for Innovation). Such a strong microbiome infrastructure as envisaged by this Community is essential to maximise the impact that European research programs have in the field of microbiome research, and to facilitate the exploitation of microbiome-based solutions in a range of settings, from clinical to industrial processes, thereby addressing key societal challenges and needs.

Data availability

No data are associated with this article.

References

1. Marchesi JR, Ravel J: **The vocabulary of microbiome research: a proposal.** *Microbiome.* 2015 Jul 30; **3**: 31.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Scientific Data.* 2016 Mar 15; **3**(1): 1–9.

3. Harrow J, Drysdale R, Smith A, *et al.*: **ELIXIR: providing a sustainable infrastructure for life science data at European scale.** *Bioinformatics.* 2021 Jun 27; **37**(16): 2506–2511.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Robertsen EM, Denise H, Mitchell A, *et al.*: **ELIXIR pilot action: Marine metagenomics – towards a domain specific set of sustainable services.** *F1000Res.* 2017 Jan 23; **6**(70): 70.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Richardson L, Allen B, Baldi G, *et al.*: **MGnify: the microbiome sequence data analysis resource in 2023.** *Nucleic Acids Res.* 2022 Dec 7; **51**(D1): D753–D759.
   Publisher Full Text

6. Agafonov A, Mattila K, Tuan CD, *et al.*: **META-pipe cloud setup and execution.** *F1000Res.* 2017 Nov 29; **6**: 2060.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Matias Rodrigues JF, Schmidt TSB, Tackmann J, *et al.*: **MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis.** *Bioinformatics.* 2017 Aug 14; **33**(23): 3808–3810.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

8. Santamaria M, Fosso B, Licciulli F, *et al.*: **ITSoneDB: a comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences.** *Nucleic Acids Res.* 2017 Sep 25; **46**(D1): D127–D132.
   Publisher Full Text

9. Nebojša T, Hervé M, Stian S-R, *et al.*: **Methods included.** *Commun. ACM.* 2022 May 20 [cited 2023 Oct 30]; **65**: 54–63.
   Publisher Full Text

10. Klemetsen T, Raknes IA, Fu J, *et al.*: **The MAR databases: development and implementation of databases specific for marine metagenomics.** *Nucleic Acids Res.* 2018 Jan 4; **46**(D1): D692–D699.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

11. Ten Hoopen P, Finn RD, Bongo LA, *et al.*: **The metagenomic data life-cycle: standards and best practices.** *Gigascience.* 2017 Aug 1; **6**(8): 1–11.
    PubMed Abstract | **Publisher Full Text**

12. Jégousse C, Vannier P, Groben R, *et al.*: **A total of 219 metagenome-assembled genomes of microorganisms from Icelandic marine waters.** *PeerJ.* 2021 Apr 2; **9**: e11112.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

13. Dávila-Ramos S, Castelán-Sánchez HG, Martínez-Ávila L, *et al.*: **A Review on Viral Metagenomics in Extreme Environments.** *Front. Microbiol.* 2019 Oct 18; **10**: 472040.
    Publisher Full Text

14. Wong HL, MacLeod FI, White RA, *et al.*: **Microbial dark matter filling the niche in hypersaline microbial mats.** *Microbiome.* 2020 Sep 16; **8**(1): 1–14.
    Publisher Full Text

15. Obiol A, Giner CR, Sánchez P, *et al.*: **A metagenomic assessment of microbial eukaryotic diversity in the global ocean.** *Mol. Ecol. Resour.* 2020 May 1; **20**(3): 718–731.
    PubMed Abstract | **Publisher Full Text**

16. Delmont TO, Gaia M, Hinsinger DD, *et al.*: **Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics.** *bioRxiv.* 2021 [cited 2023 Oct 30]; p. 2020.10.15.341214.
    Publisher Full Text

17. Tagirdzhanova G, Saary P, Cameron ES, *et al.*: **Evidence for a core set of microbial lichen symbionts from a global survey of metagenomes.** *bioRxiv.* 2023 [cited 2023 Oct 30]; p. 2023.02.02.524463.
    Publisher Full Text

18. Alberdi A, Andersen SB, Limborg MT, *et al.*: **Disentangling host-microbiota complexity through hologenomics.** *Nat. Rev. Genet.* 2022 May; **23**(5): 281–297.
    PubMed Abstract | **Publisher Full Text**

19. Nielsen HB, Almeida M, Juncker AS, *et al.*: **Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes.** *Nat. Biotechnol.* 2014 Aug; **32**(8): 822–828.
    PubMed Abstract | **Publisher Full Text**

20. Lu J, Rincon N, Wood DE, *et al.*: **Metagenome analysis using the Kraken software suite.** *Nat. Protoc.* 2022 Dec; **17**(12): 2815–2839.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

21. Beghini F, McIver LJ, Blanco-Míguez A, *et al.*: **Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.** *elife.* 2021 May 4; **10**.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

22. Ruscheweyh HJ, Milanese A, Paoli L, *et al.*: **Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments.** *Microbiome.* 2022 Dec 5; **10**(1): 212.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

23. Martin S, Heavens D, Lan Y, *et al.*: **Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples.** *Genome Biol.* 2022 Jan 24; **23**(1): 1–27.
    Publisher Full Text

24. Nelson MT, Pope CE, Marsh RL, *et al.*: **Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles.** *Cell Rep.* 2019 Feb 19; **26**(8): 2227–40.e5.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

25. Lin Z, Akin H, Rao R, *et al*.: **Evolutionary-scale prediction of atomic-level protein structure with a language model.** *Science.* 2023 Mar 17; **379**(6637): 1123–1130.
PubMed Abstract | **Publisher Full Text**

26. Bowers RM, Kyrpides NC, Stepanauskas R, *et al*.: **Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.** *Nat. Biotechnol.* 2017 Aug 1; **35**(8): 725–731.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

27. Arikawa K, Ide K, Kogawa M, *et al*.: **Recovery of strain-resolved genomes from human microbiome through an integration framework of single-cell genomics and metagenomics.** *Microbiome.* 2021 Oct 12; **9**(1): 1–16.
Publisher Full Text

28. Ghaddar B, Biswas A, Harris C, *et al*.: **Tumor microbiome links cellular programs and immunity in pancreatic cancer.** *Cancer Cell.* 2022 Oct 10; **40**(10): 1240–1253.e5.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Balech B, Brennan L, de Santa Pau EC, *et al*.: **The future of food and nutrition in ELIXIR.** *F1000Res.* 2022 Aug 25; **11**(978): 978.
Publisher Full Text

30. Heberling JM, Miller JT, Noesgaard D, *et al*.: **Data integration enables global biodiversity synthesis.** *Proc. Natl. Acad. Sci. U. S. A.* 2021 Feb 9; **118**(6).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

31. Vizcaíno JA, Walzer M, Jiménez RC, *et al*.: **A community proposal to integrate proteomics activities in ELIXIR.** *F1000Res.* 2017 Jun 13; **6**: 875.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

32. Bansal P, Morgat A, Axelsen KB, *et al*.: **Rhea, the reaction knowledgebase in 2022.** *Nucleic Acids Res.* 2022 Jan 7; **50**(D1): D693–D700.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

33. Jumper J, Evans R, Pritzel A, *et al*.: **Highly accurate protein structure prediction with AlphaFold.** *Nature.* 2021 Aug; **596**(7873): 583–589.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

34. Varadi M, Anyango S, Deshpande M, *et al*.: **AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.** *Nucleic Acids Res.* 2021 Nov 17; **50**(D1): D439–D444.
Publisher Full Text

35. Van Den Bossche T, Arntzen MØ, Becher D, *et al*.: **The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes.** *Microbiome.* 2021 Dec 20; **9**(1): 243.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

36. Yoshida S, Hiraga K, Takehana T, *et al*.: **Response to Comment on "A bacterium that degrades and assimilates poly (ethylene terephthalate).".** *Science.* 2016 Aug 19; **353**(6301): 759.
PubMed Abstract | **Publisher Full Text**

37. Gurbich TA, Almeida A, Beracochea M, *et al*.: **MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues.** *J. Mol. Biol.* 2023 Jul 15; **435**(14): 168016.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

38. Claeys T, Van Den Bossche T, Perez-Riverol Y, *et al*.: **lesSDRF is more: maximizing the value of proteomics data through streamlined metadata annotation.** *Nat. Commun.* 2023 Oct 24; **14**(1): 1–4.
Publisher Full Text

39. Dai C, Füllgrabe A, Pfeuffer J, *et al*.: **A proteomics sample metadata representation for multiomics integration and big data analysis.** *Nat. Commun.* 2021 Oct 6; **12**(1): 1–8.
Publisher Full Text

40. Haug K, Cochrane K, Nainala VC, *et al*.: **MetaboLights: a resource evolving in response to the needs of its scientific community.** *Nucleic Acids Res.* 2019 Nov 6; **48**(D1): D440–D444.
Publisher Full Text

41. The Europe PMC Consortium: **Europe PMC: a full-text literature database for the life sciences and platform for innovation.** *Nucleic Acids Res.* 2014 Nov 6; **43**(D1): D1042–D1048.
Publisher Full Text

42. Nassar M, Rogers AB, Talo' F, *et al*.: **A machine learning framework for discovery and enrichment of metagenomics metadata from open access publications.** *Gigascience.* 2022 Aug 11; **11**.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

43. Zafeiropoulos H, Paragkamian S, Ninidakis S, *et al*.: **PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types.** *Microorganisms.* 2022 Jan 26; **10**(2).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

44. Grueening B, Sallou O, Moreno P, *et al*.: **Recommendations for the packaging and containerizing of bioinformatics software.** *F1000Research.* 2018: 7.

45. Ison J, Kalas M, Jonassen I, *et al*.: **EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats.** *Bioinformatics.* 2013 May 15; **29**(10): 1325–1332.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

46. Ison J, Rapacki K, Ménager H, *et al*.: **Tools and data services registry: a community effort to document bioinformatics resources.** *Nucleic Acids Res.* 2015; **44**: D38–D47.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

47. Goble C, Soiland-Reyes S, Bacall F, *et al*.: **Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory.** *Zenodo.* 2021.
Reference Source

48. Lindgreen S, Adair KL, Gardner PP: **An evaluation of the accuracy and speed of metagenome analysis tools.** *Sci. Rep.* 2016 Jan 18; **6**(1): 1–14.
Publisher Full Text

49. Wu Z, Wang Y, Zeng J, *et al*.: **Constructing metagenome-assembled genomes for almost all components in a real bacterial consortium for binning benchmarking.** *BMC Genomics.* 2022 Nov 10; **23**(1): 1–19.
Publisher Full Text

50. Poussin C, Khachatryan L, Sierro N, *et al*.: **Crowdsourced benchmarking of taxonomic metagenome profilers: lessons learned from the sbv IMPROVER Microbiomics challenge.** *BMC Genomics.* 2022 Aug 30; **23**(1): 1–19.
Publisher Full Text

51. Almeida A, Mitchell AL, Tarkowska A, *et al*.: **Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments.** *Gigascience.* 2018 May 11; **7**(5): giy054.
Publisher Full Text

52. O'Sullivan DM, Doyle RM, Temisak S, *et al*.: **An inter-laboratory study to investigate the impact of the bioinformatics component on microbiome analysis using mock communities.** *Sci. Rep.* 2021 May 19; **11**(1): 10590.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

53. Sczyrba A, Hofmann P, Belmann P, *et al*.: **Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software.** *Nat. Methods.* 2017 Nov; **14**(11): 1063–1071.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

54. Fritz A, Hofmann P, Majda S, *et al*.: **CAMISIM: simulating metagenomes and microbial communities.** *Microbiome.* 2019 Feb 8; **7**(1): 17.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

55. Meyer F, Hofmann P, Belmann P, *et al*.: **AMBER: Assessment of Metagenome BinnERs.** *Gigascience.* 2018 Jun 1; **7**(6).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

56. Meyer F, Fritz A, Deng ZL, *et al*.: **Critical Assessment of Metagenome Interpretation: the second round of challenges.** *Nat. Methods.* 2022 Apr; **19**(4): 429–440.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

57. Perkel JM: **Workflow systems turn raw data into scientific knowledge.** *Nature.* 2019 Sep; **573**(7772): 149–150.
PubMed Abstract | **Publisher Full Text**

58. Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory: [cited 2023 Oct 30].
Reference Source

59. Zafeiropoulos H, Beracochea M, Ninidakis S, *et al*.: **metaGOflow: a workflow for the analysis of marine Genomic Observatories shotgun metagenomics data.** *Gigascience.* 2022 Dec 28; **12**.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

60. Soiland-Reyes S, Sefton P, Crosas M, *et al*.: **Packaging research artefacts with RO-Crate.** *Data Sci.* 2022; **5**(2): 97–138.
Publisher Full Text

61. Beard N, Bacall F, Nenadic A, *et al*.: **TeSS: A Platform for Discovering Life-Science Training Opportunities.** *Bioinformatics.* 2020; **36**(10): 3290–3291.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

62. Field D, Amaral-Zettler L, Cochrane G, *et al*.: **The Genomic Standards Consortium.** *PLoS Biol.* 2011 Jun; **9**(6): e1001088.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

63. Yilmaz P, Kottmann R, Field D, *et al*.: **Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.** *Nat. Biotechnol.* 2011 May 6; **29**(5): 415–420.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

64. McDonald D, Clemente JC, Kuczynski J, *et al*.: **The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome.** *Gigascience.* 2012 Jul 12; **1**(1): 2047–217X – 1–7.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

65. Meisner A, Kostic T, Vernooij M, *et al*.: **The global microbiome research landscape: mapping of research, infrastructures,**

policies and institutions in 2021. *MicrobiomeSupport Consortium.* 2022.

66. Van Den Bossche T, Kunath BJ, Schallert K, *et al.*: **Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows.** *Nat. Commun.* 2021 Dec 15; **12**(1): 1–15.
Publisher Full Text

67. Eloe-Fadrosh EA, Ahmed F, Anubhav A, *et al.*: **The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource.** *Nucleic Acids Res.* 2021 Oct 30; **50**(D1): D828–D836.

68. Parks DH, Imelfort M, Skennerton CT, *et al.*: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.** *Genome Res.* 2015 Jul; **25**(7): 1043–1055.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

69. Parks DH, Chuvochina M, Rinke C, *et al.*: **GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy.** *Nucleic Acids Res.* 2021 Sep 14; **50**(D1): D785–D794.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

70. Chaumeil PA, Mussig AJ, Hugenholtz P, *et al.*: **GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.** *Bioinformatics.* 2019 Nov 15; **36**(6): 1925–1927.
PubMed Abstract | **Publisher Full Text**

71. Keegan KP, Glass EM, Meyer F: **MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function.** *Methods Mol. Biol.* 2016; **1399**: 207–233.
PubMed Abstract | **Publisher Full Text**

72. Chen IMA, Chu K, Palaniappan K, *et al.*: **The IMG/M data management and analysis system v.7: content updates and new features.** *Nucleic Acids Res.* 2023 Jan 6; **51**(D1): D723–D732.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

73. Camargo AP, Nayfach S, Chen IMA, *et al.*: **IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata.** *Nucleic Acids Res.* 2023 Jan 6; **51**(D1): D733–D743.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

74. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, *et al.*: **The European Nucleotide Archive in 2019.** *Nucleic Acids Res.* 2020 Jan 8; **48**(D1): D70–D76.
PubMed Abstract | **Publisher Full Text**

75. **CAMI II: identifying best practices and issues for metagenomics software.** *Nat. Methods.* 2022 Apr; **19**(4): 412–413.
PubMed Abstract | **Publisher Full Text**

76. Sommers P, Chatterjee A, Varsani A, *et al.*: **Integrating Viral Metagenomics into an Ecological Framework.** *Annu Rev Virol.* 2021 Sep 29; **8**(1): 133–158.
PubMed Abstract | **Publisher Full Text**

77. Roux S, Camargo AP, Coutinho FH, *et al.*: **iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria.** *PLoS Biol.* 2023 Apr; **21**(4): e3002083.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

78. Camargo AP, Roux S, Schulz F, *et al.*: **Identification of mobile genetic elements with geNomad.** *Nat. Biotechnol.* 2023 Sep 21: 1–10.
Publisher Full Text

79. **Genome-resolving metagenomics reveals wild western capercaillies (Tetrao urogallus) as avian hosts for antibiotic-resistance bacteria and their interactions with the gut-virome community.** *Microbiol. Res.* 2023 Jun 1; **271**: 127372.
PubMed Abstract | **Publisher Full Text**

80. Fredslund F, Borchert MS, Poulsen JCN, *et al.*: **Structure of a hyperthermostable carbonic anhydrase identified from an active hydrothermal vent chimney.** *Enzym. Microb. Technol.* 2018 Jul; **114**: 48–54.
PubMed Abstract | **Publisher Full Text**

81. Schiebenhoefer H, Van Den Bossche T, Fuchs S, *et al.*: **Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis.** *Expert Rev. Proteomics.* 2019 May; **16**(5): 375–390.
PubMed Abstract | **Publisher Full Text**

82. Felden J, Möller L, Schindler U, *et al.*: **PANGAEA - Data Publisher for Earth & Environmental Science.** *Scientific Data.* 2023 Jun 2; **10**(1): 1–9.
Publisher Full Text

83. Argelaguet R, Velten B, Arnol D, *et al.*: **Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets.** *Mol. Syst. Biol.* 2018 Jun 1; **14**(6): e8124.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000 Research