OPINION ARTICLE

REVISED Establishing the ELIXIR Microbiome Community

[version 2; peer review: 2 approved, 1 approved with reservations]

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], Teresa Nogueira[14,15], Evangelos Pafilis[16], Graziano Pesole[2,8], Lorna Richardson[1], Monica Santamaria[17], Nikolaos Strepis[18], Tim Van Den Bossche[19,20], Juan Antonio Vizcaíno[1], Haris Zafeiropoulos[16], Nils P. Willassen[21], Eric Pelletier[12,22], Bérénice Batut[13,23]

[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 Université, 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]Metabolic Genomics, Genoscope, Institut François-Jacob / CEA / CNRS / Université Evry / Université Paris-Saclay, Evry, France
[13]IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800, Villejuif, France
[14]INIAV—National Institute for Agrarian and Veterinary Research, 4485-655, Vairão, Portugal
[15]CE3c - Centre for Ecology, Evolution and Environmental Changes & CHANGE - Global Change and Sustainability Institute, Faculdade de Ciências da Universidade de Lisboa, 1749-016, Lisboa, Portugal
[16]Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Greece
[17]Department of Soil, Plant and Food Sciences (Di.S.S.P.A.), University of Bari, Bari, Italy
[18]Department of Pathology and Clinical Bioinformatics, Erasmus MC Cancer Institute, Erasmus MC, Rotterdam, The Netherlands
[19]Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent, Belgium
[20]VIB, UGent Center for Medical Biotechnology, Ghent, Belgium
[21]UiT The Arctic University of Norway, Tromsø, Norway
[22]Research Federation for the study of Global Ocean Systems Ecology and Evolution, CNRS, Paris, France
[23]Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Freiburg, Germany

v2 **First published:** 08 Jan 2024, **13**(ELIXIR):50
https://doi.org/10.12688/f1000research.144515.1
Latest published: 08 Sep 2025, **13**(ELIXIR):50
https://doi.org/10.12688/f1000research.144515.2

Abstract

Microbiome research has grown substantially over the past decade in terms of the range of biomes sampled, identified taxa, and the

Open Peer Review

Approval Status ✓ ✓ ?

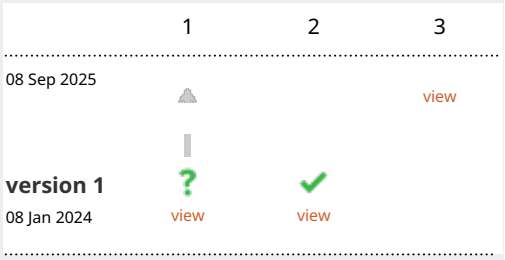
	1	2	3
version 2 (revision)	✓ view		?

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 identifying solutions to the computational overheads of performing such analyses. Nevertheless, the methods used and the challenges faced are not confined to marine microbiome studies, but are broadly applicable to other biomes. Thus, expanding this Marine Metagenomics Community to a more inclusive ELIXIR Microbiome Community will enable it to encompass a broader 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.

This article is included in the EMBL-EBI collection.



1. **Charlie Pauvert**, University Hospital of RWTH, Aachen, Germany
2. **Almut Heinken**, University of Lorraine, Lorraine, France
3. **Lauren Lui**, E O Lawrence Berkeley National Laboratory, Berkeley, USA

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

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; **Nogueira T**: 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; **Strepis N**: Writing – Review & Editing; **Van Den Bossche T**: Writing – Original Draft Preparation, Writing – Review & Editing; **Vizcaíno JA**: Writing – Original Draft Preparation, Writing – Review & Editing; **Zafeiropoulos H**: Writing – Original Draft Preparation, Writing – Review & Editing; **Willassen NP**: 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

Competing interests: No competing interests were disclosed.

Grant information: CJC received Portuguese national funds from the Foundation for Science and Technology (FCT) through projects UIDB/04326/2020, UIDP/04326/2020, and LA/P/0101/2020. T.V.D.B. acknowledges funding from the Research Foundation Flanders (FWO) [1286824N]. GP acknowledges funding from MUR (Italy), CnrBiomics (grant number PIR01_00017) and ELIXIRxNextGenIT (grant number IR0000010). JUV received the C19/BM/13684739 grant, funded by National Research Fund Luxembourg (FNR). VK was supported by a Biotechnology and Biological Sciences Research Council [BB/V01868X/1]. LR and RDF were supported by EMBL core funds.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Finn RD, Balech B, Burgin J *et al.* **Establishing the ELIXIR Microbiome Community [version 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2025, **13**(ELIXIR):50 https://doi.org/10.12688/f1000research.144515.2

First published: 08 Jan 2024, **13**(ELIXIR):50 https://doi.org/10.12688/f1000research.144515.1

REVISED **Amendments from Version 1**

In response to insightful feedback from our reviewers, this revised version of our article has undergone significant updates and refinements to enhance clarity, accuracy, and comprehensiveness. We have meticulously addressed the comments provided, ensuring that each section of the manuscript is thoroughly revised to reflect the latest developments and insights in the field.

One of the most notable changes in this version is the substantial reworking of Tables 2 and 3. Table 2, which provides an overview of existing and planned interactions with ELIXIR Communities, has been expanded and updated to include new collaborations and initiatives that have emerged since the previous publication. This table now offers a more detailed and current perspective on how our work integrates with and supports the broader ELIXIR infrastructure.

Similarly, Table 3, which describes a selection of current European national and pan-European efforts aimed at microbiome research, has been comprehensively revised. We have added new entries and updated existing ones to reflect the latest advancements and their relevance to the ELIXIR Microbiome Community. This table now provides a more accurate and up-to-date snapshot of the European microbiome research landscape.

In addition to these updates, the text throughout the article has been carefully revised to improve readability and coherence. We have also included new data and references to recent studies to ensure that our discussion is grounded in the most current research.

Overall, this revised version aims to provide a more accurate, comprehensive, and timely overview of our work and its implications for the ELIXIR Microbiome Community and the broader field of microbiome research.

Any further responses from the reviewers can be found at the end of the article

1. Introduction

The term "microbiome" is a description of an entire habitat that encompasses all the microbes (bacteria, archaea, eukaryotes, and viruses), their composition (genomes, proteins and various molecules they produce), and the environment they are found in.[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. Furthermore, there can be complex compositional modulation of microbiomes. For example, viruses that infect bacteria are found ubiquitously in all environments and play critical roles in community dynamics. 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 close to a natural composition are key current research questions. Given that most academic institutions have access to dedicated sequencing facilities or equivalent commercial facilities, coupled with diminishing costs of DNA sequencing and other 'omics technologies, it is relatively easy to generate large datasets. Therefore, there are now millions of microbiome-derived sequence datasets, many of which are large (gigabytes to terabytes) and complex (thousands of related and/or diverse samples), but it can require significant computational resources to store and analyse the data. Additionally, datasets from other 'omics techniques such as metaproteomics and metabolomics 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 reusable, i.e. FAIR.[2]

Such steps are vital to ensure reuse of data and scientific reproducibility. For example, when wishing to contextualise the results between 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 (ASVs) can only be compared to those with the same amplified region). Similarly, ensuring that taxonomic and functional assertions are placed in the context of the original sample/sequencing effort and associated contextual metadata is crucial for understanding compositional microbiome changes, such as those in health and disease or longitudinal datasets. Ensuring data is FAIR is typically necessary to comply with most scientific funding sources open access data requirements. Finally, describing how a dataset has been produced allows verification and replication of scientific experiments.

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. During the establishment of ELIXIR, 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 biome. Furthermore, there is the need to extend the activities of the Marine Metagenomics Community to integrate expertise and knowledge about other 'omics technologies, such as metatranscriptomics, metaproteomics and metabolomics, which are increasingly used in microbiome studies. Thus, this white paper outlines some of the historical aspects of the Marine Metagenomics Community and the aims of the broader community, especially in the context of the other ELIXIR Communities and infrastructure platforms.

2. 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,[7] ITSOneDB[8]), the formal description of the MGnify pipeline using the common workflow language (CWL[9]) to promote interoperability, the establishment of marine metagenomics data (e.g. Marine Metagenomics Portal, MARdb,[10] METDB,[11] and the Ocean Gene Atlas[12]) 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.[13] Capacity building has also been an important activity since the establishment of the Marine Metagenomics 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.[14–18] 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.[19] 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 host, such as genome, single nucleotide polymorphisms and transcriptomic data. The collective data facilitate a hologenomic approach[20] 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.

2.1 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), as exemplified by the thousands of mislabeled datasets in International Nucleotide Sequence Database Collaboration (INSDC). 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).[21] Equally applicable is the analysis of unassembled raw-read data sets that can be used for taxonomic classification

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 and their interactions found in a microbial community

(e.g. Kraken,[22] MetaPhlan,[23] mOTUs[24]) and functional profiling approaches that are especially effective when extensive reference databases are available. Sequencing technologies such as long-read sequencing methodologies and the associated adaptive sequencing techniques,[25] together with changing protocols such as host material depletion protocols (e.g. [26]), 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 metagenomic 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 and metabolomics are technologies that are increasingly being used in microbiome research, involving the study of proteins expressed or small molecules produced by microbial communities in a given environment, which require quite different methodologies for their analysis.

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. The Microbiome 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 the Microbiome Community is to address the provision of infrastructures that are sufficiently adaptable to permit the most appropriate informatics analysis, depending on the environment sampled and the experiments conducted. Finally, the ELIXIR Microbiome Community will gather a 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.

2.2 The context within ELIXIR

Given the breadth of the 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.

2.2.1 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.[27] Thus, in the case of the Food

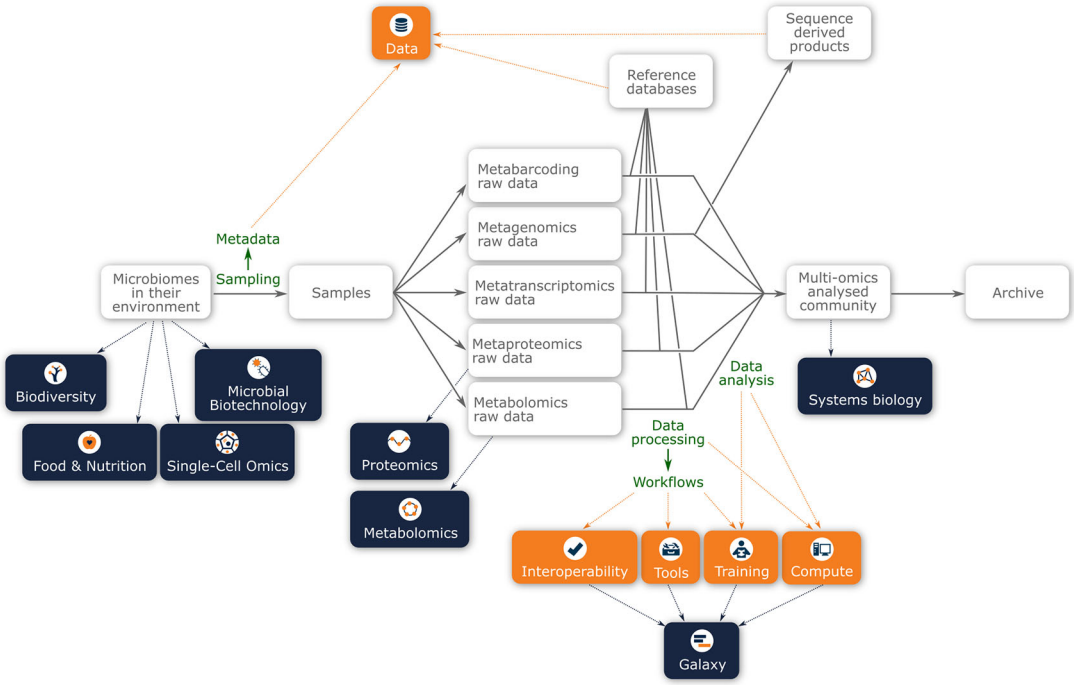


Figure 1. A schematic of how a microbiome sample (i.e. community in the environment) may be analysed using different 'omics approaches, with the main steps indicated in green. Underpinning these analyses will be the metagenomic and metatranscriptomic data, which will be used as a framework for the metaproteomic and metabolomic interpretation. Highlights in this figure are connections with the ELIXIR platforms (orange boxes) and other ELIXIR communities (dark blue boxes).

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

ELIXIR Community	Existing and planned interaction(s)
3D BioInfo	Improve the organisation, quality control and presentations of protein models (e.g. ESMAtlas[28]) for proteins predicted from metagenomic and metatranscriptomic assemblies. Improve functional annotations through structure-function-sequence relationships.
Biodiversity	Connect biodiversity/observation resources with 'omics data/analysis. Identify overlap between analysis pipelines (e.g. barcodes, genome annotations) and promote best practices.
Federated human data	Evaluate the landscape concerning human microbiome and national legislation concerning data sharing. If appropriate, investigate solutions from the Federated human data Community for sharing sensitive data.
Food & Nutrition	Share metagenomics workflows to improve our understanding of the role of the gut microbiome in unlocking nutrients in food.
Galaxy	In collaboration with the Galaxy Community, continue to tailor and expand tools, workflows and training materials applicable to the Microbiome Community, directed by the needs identified by an ongoing evaluation study.
Metabolomics	Develop methods and tools to connect microbiome sequence data (metagenomics and metatranscriptomics) to link functions to metabolites.
Microbial Biotechnology	Improve the identification of valuable enzymatic activities from environmental metagenomics data to identify bioactives (e.g. enzyme, small molecule) of interest for: bioeconomy; applications in food preservation; agriculture; chemistry; or medicine.
Plant Science	Develop a greater understanding of the needs of the Plant Science Community for microbiome-based solutions to improve plant resilience to pathogens, as well as understand how plants maintain their microbial communities across generations, and if so, potential mechanisms for doing so.

Table 2. *Continued*

ELIXIR Community	Existing and planned interaction(s)
Proteomics	Enable the production of tailored reference databases (e.g. biome and/or other contextual metadata) for interpretation of (meta-)proteomics MS2 data. Develop methods to enhance the integration of metagenomic, metatranscriptomics and metaproteomics results.
Single Cell Omics	Explore areas of overlap in data standards[29] and annotation pipelines concerning single amplified genomes (SAGs). Increase knowledge within the Microbiome Community concerning spatial single cell data with respect to improving the quality of MAGs/SAGs[30] and the identification of microbes in tissues and tumours.[31]
Systems Biology	Empower a better integration of multi-omics datasets to describe and understand how different community members interoperate to achieve processes. More specifically, ensure that different multi-omics data types from the same sample are appropriately connected across different archive databases, and improve methods for linking metabolomics data to sequence (protein and nucleotide).

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 Microbiome Community are already engaged with the Food and 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[32] routinely incorporate data from both MGnify and INSDC. Similarly, many of the biodiversity approaches use barcoding methods for studying environmental DNA (eDNA). While this can overlap 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.

With the growing number of multi-omics datasets, establishing strong ties with the ELIXIR Metabolomics and Proteomics communities[33] will be essential for understanding how metagenomic and metatranscriptomic data may be utilised by these Communities (e.g. the production of reference databases for the interpretation of the metaproteomics), 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.[34] 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[28,35] and data resources[36] 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.

Furthermore, microbiome research has many translational aspects, ranging from the discovery of biomarkers associated with health (of organisms or environments) 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 (commonly known as PET).[37] While metagenomic assembly and analysis is providing a rich source of potential 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.

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 global societal challenges within the One Health concept defined by the World Health Organization, such as food safety, pathogen

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

Initiative	Reach	Aims and relevance
Mutualised Digital Spaces For Life Sciences (MuDIS4LS)	National (FR)	To develop a framework that will connect national and regional data centres to enable the control of biological data from their origin (data-producing national infrastructures) to their public release, while ensuring data security during the intermediate phases of analysis and exploitation. *Relevance*: Guide data management best practices, especially when dealing with sensitive microbiome related data.
Secured computing spaces for the data access and analysis project of the France 2030 programme « Food Systems, Microbiome and Health »(Cloud4SAMS)	National (FR)	To deploy a distributed digital infrastructure enabling researchers to exploit microbiome and health data in a secure computing environment, collating software tools and workflows for processing these data, computing and storage platforms suitable for processing microbiome data and matching them with health data, while respecting data access rules. *Relevance*: This project will define deployment recipes describing all the procedures to instantiate a virtual machine in a secure cloud, install software and transfer datasets. Knowledge developed will guide the deploying of similar distributed infrastructures within ELIXIR.
National Research Center in Bioinformatics for Omics Sciences (CNRBiOmics)	National (IT)	To enhance the ELIXIR Italian node infrastructure, through the establishment of a "centre of excellence" for multi-omics data production, management, and analysis. In addition, establish a higher education training platform to develop skills required to use the infrastructure. *Relevance*: Microbiome multi-omics data produced by the infrastructure will provide example use cases, as well as being a test bed for multi-omics data integration solutions developed by the community.
Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics (ELIXIRxNextGenIT)	National (IT)	To consolidate the ELIXIR-IT infrastructure for omics and bioinformatics, focused on data production, computational analysis, facilities improvement and training, with a view to strengthening the national ELIXIR infrastructure. *Relevance*: Overlaps with training and sharing of microbiome related pipelines.
NFDI4Microbiota	National (DE)	To: (i) promote FAIR principles in the microbiological community; (ii) provide a comprehensive training program; (iii) enhance data resources for microbiology community; (iv) support high-quality research data management; (v) increase data value 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. *Relevance*: This national programme shares many of the objectives of the Microbiome Community, so it will be important to synergise activities.

Table 3. *Continued*

Initiative	Reach	Aims and relevance
European Reference Genomes Atlas (ERGA)	European	To generate eukaryotic reference genomes of European species and create a powerful resource for the understanding of biodiversity. *Relevance*: This collection will also include microbial eukaryotes that can be used to enhance eukaryote genome analysis in microbiome research.
European e-Science Infrastructure for biodiversity and ecosystem research (LifeWatch ERIC)	European	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. *Relevance*: Metagenomics data will be used to develop ecological models, while reference genome data will improve analysis pipelines.
Metaproteomics Initiative	European/ International	To promote dissemination of metaproteomics fundamentals, advancements, and applications through collaborative networking in microbiome research. *Relevance:* The central information hub and open meeting place will allow members of the Microbiome Community to interact with metaproteomics experts. Will help the Microbiome Community aim to standardise and methodologies in this field.
microGalaxy	European/ International	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. *Relevance*: Usable and standardised workflows for the Microbiome Community.

tracking, climate changes, antimicrobial resistance (and new therapeutics) and pandemic preparedness, in which microbiome research plays a critical role. Yet each one of these areas is too complex to be tackled individually and therefore requires the collective outputs from more than one ELIXIR Community, and reach far beyond informatics research (Table 3).

2.2.2 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.

2.2.2.1 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 INSDC, especially the European Nucleotide Archive (ENA) in the context of ELIXIR. 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.[38] 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.[39,40] 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[41] 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[42] and other CDR databases. This is critical for the Microbiome Community as additional contextual metadata can often be found in the literature,[43,44] providing crucial overarching context to the experiment, which can be important for re-analyses or 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 ELIXIR Research Data Management Kit (RDMKit) towards a more general Microbiome domain.

2.2.2.2 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[45] to promote the packaging, containerisation and deployment of tools relevant to microbiome research.

In order to make tools findable by the end users, the Microbiome Community will work on improving their annotation by (i) expanding the EDAM ontology[46] to include microbiome-specific keywords, (ii) performing periodic reviews of tools and their associated annotations in the bio.tools[47] 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,[48] or as workflow descriptions (e.g. Snakemake,[49] CWL,[9] Nextflow[50]), which can be readily combined to make new annotation workflows. Additionally, the Microbiome 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[51] for easy discovery, re-use and assessment; (iii) making them available for the Microbiome 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 current 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.

A key part of understanding the applicability of a tool/workflow is its benchmarking, and has been an aim of the Tools platform. 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. [47,51–54]), with the Critical Assessment of Microbiome Interpretations (CAMI) having become an internationally recognised benchmarking effort.[55–58] 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 wider microbiome research community needs, and feed these requirements 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.

2.2.2.3 Compute

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.

2.2.2.4 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,[59] which also allowed MGnify and MG-RAST to exchange pipelines 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.[60] Since then, MGnify has published their workflows in WorkflowHub, 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,[61] to be used by the Marine Genomic Observatories. Moreover, this work also employed Research Object Crate (RO-crate)[62] 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 Global Alliance for Genomics and Health (GA4GH) standards, and is in the process of applying to be an ELIXIR 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 Microbiome 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 used in metagenomics-centred Microbiome data, such as the PRIDE database in the case of metaproteomics data.

2.2.2.5 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 bioinformatics. 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, and sets the need for well defined and consistent training modules dedicated to computational approaches to microbiome analysis.

Platforms such as MGnify support large-scale services for most 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[29]; (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, digital 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 strategies. 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.

3. 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 Genomics Standards Consortium (GSC[63]). The GSC has become critical for establishing many of the standards that underpin genomic research, and more recently metabolomic. Examples of GSC established standards that are particularly pertinent to the microbiome domain include: minimal information about any sequence (MIxS[64]), the Biological Observation Matrix (BIOM) format[65]; and the Minimum Information about a Metagenome-Assembled Genome (MIMAG[29]).

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 coordination and support action (CSA) program aimed at improving microbiome research and innovation, highlighted in their final report[66] 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. Since the end of the CSA project, this MicrobiomeSupport has transitioned to the MicrobiomeSupport Association to continue the engagement activities within the microbiome research community and will provide a key dissemination route for the ELIXIR Microbiome Community outcomes.

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.[67,68] 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.[67]

Finally, the National Microbiome Data Collaborative (NMDC),[69] 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.

4. 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,[70] the Genome Taxonomy Database (GTDB) and the associated GTDB toolkit.[71,72] 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[59] 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[73] 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,[74] 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. Recently it has also produced IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata.[75] 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.

5. 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

Table 4. Objectives of the ELIXIR Microbiome Community.

Area	Objective
Near-term (2 years timeframe)	
Community Expansion	Survey the 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 data analysis workflows
	Teach advanced containerisation and cloud deployment
Co-ordinate	Increase rates of data archival deposition, with rich contextual metadata. Establish a mapping between biome and checklists
	Promote data analysis through the use of ELIXIR services
	Share 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	Use ELIXIR and Node forums to understand pharmaceutical and biotechnological demands and current limitations impacting this sector.
Longer-term (~3-5 years)	
Training	Design 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 find ability
Federated data analysis	Enable the execution of MGnify pipelines in Galaxy and/or other data 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	Establish new standards for microbiome research, particularly with respect to data analysis reporting and contextual metadata reporting in conjunction with GSC
	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 the Microbiome Community demand and capacity
	Establish 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 the 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

development of new experimental methods, from sampling 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 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 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 environmental genomes (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,[76] 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 CAMI[55,77] 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[71] and the associated GTDB-tk[72] 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 of taxonomy 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.[78–81] Viral genomes are incredibly diverse in terms of composition and organisation. 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, and how best to overcome the current fragmented organisation of viral datasets.

The increase in metagenomic assemblies has resulted in a parallel increase in the number of predicted 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[82] 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 sequence based methods. The Microbiome 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 or adequate representatives 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 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 Microbiome 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 Microbiome Community, we will promote and develop standards regarding the analysis provenance, 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 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 extensions to existing databases or development of new ones, 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 One Health 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.[83] 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). Other challenges in metaproteomics includes how to report proteins which cannot be explained by a corresponding metagenomic study, so called unidentified proteins of unknown function, methods for protein inference and quantification that are comparable across studies, the dynamic range of protein sizes analysed, as well as how researchers handle

non-specific peptide-spectrum matches, compared to specific matches. Addressing these challenges 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[84]) measurements, meteorological, image data and even acoustics. While methods are already emerging for the integration of 'omics datatypes (e.g. MOFA,[85] MIA (https://github.com/microbiome/mia)), integration of these additional non-omics data types will enable a broader understanding 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.

6. 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. The aim is to be biome-agnostic and, 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 Marine Metagenomics 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 Microbiome 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 the Microbiome 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 is associated with this article.

Acknowledgments

We would like to thank the reviewers of the manuscript for their insightful comments, particularly Reviewer 1 for providing a very thorough review, with many constructive comments that have been incorporated into the final manuscript.

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. Crusoe MR, Abeln S, Iosup A, *et al.*: **Methods included: standardizing computational reuse and portability with the Common Workflow Language.** *Commun. ACM.* 2022 May 20; **65**(6): 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. Niang G, Hoebeke M, Meng A, *et al.*: **METdb: A Genomic Reference Database For Marine Species.** *F1000Res.* 2020 Jun 6; **9**(ELIXIR): 564 (poster).
Publisher Full Text

12. Vernette C, Lecubin J, Sánchez P, *et al.*: **The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes.** *Nucleic Acids Res.* 2022 Jul 5; **50**(W1): W516–W526.
PubMed Abstract | Publisher Full Text | Free Full Text

13. 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

14. 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

15. 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

16. 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

17. 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

18. Delmont TO, Gaia M, Hinsinger DD, *et al.*: **Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics.** *bioRxiv.* 2020.10.15.341214.
Publisher Full Text

19. Tagirdzhanova G, Saary P, Cameron ES, *et al.*: **Evidence for a core set of microbial lichen symbionts from a global survey of metagenomes.** *bioRxiv.* 2023.02.02.524463.
Publisher Full Text

20. 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

21. 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

22. 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

23. 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

24. 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

25. 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

26. 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

27. Balech B, Brennan L, Carrillo de Santa Pau E, *et al.*: **The future of food and nutrition in ELIXIR.** *F1000Res.* 2022 **11**(ELIXIR): 978.
Publisher Full Text

28. 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

29. Beard N, Bacall F, Nenadic A, Thurston M: **Carole A Goble, Susanna-Assunta Sansone, Teresa K Attwood, TeSS: a platform for discovering life-science training opportunities.** *Bioinformatics.* 2020; **36**(10): 3290–3291.
PubMed Abstract | Publisher Full Text | Free Full Text

30. 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): 202.
PubMed Abstract | Publisher Full Text | Free Full Text

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

32. Heberling JM, Miller JT, Noesgaard D, *et al.*: **Data integration enables global biodiversity synthesis.** *PNAS.* 2021 Feb 9; **118**(6): e2018093118.
PubMed Abstract | Publisher Full Text | Free Full Text

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

34. 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

35. 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

36. 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

37. 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

38. 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

39. 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

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

41. 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

42. 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

43. 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

44. 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

45. Gruening B, Sallou O, Moreno P, *et al.*: **Recommendations for the packaging and containerizing of bioinformatics software.** *F1000Research.* 2018; **7**.
Publisher Full Text

46. 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

47. 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.
Publisher Full Text

48. The Galaxy Community: **The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update.** *Nucleic Acids Res.* 2024; **52**(W1): W83–W94.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Mölder F, Jablonski KP, Letcher B, *et al.*: **Sustainable data analysis with Snakemake.** *F1000Res.* 2021; **10**: 33.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat. Biotechnol.* 2017 Apr; **35**(4): 316–319.
PubMed Abstract | Publisher Full Text

51. Gustafsson OJR , *et al.*: **WorkflowHub: a registry for computational workflows.** *Sci. Data* 2025; **12**: 837.
PubMed Abstract | Publisher Full Text | Free Full Text

52. 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

53. 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

54. 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): 624.
PubMed Abstract | Publisher Full Text | Free Full Text

55. 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

56. 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

57. 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

58. 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

59. 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

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

61. 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

62. 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

63. 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

64. 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

65. 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

66. 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.

67. 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

68. 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

69. 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.

70. 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

71. 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

72. 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

73. 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

74. 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

75. Camargo AP, Call L, Roux S, *et al.*: **IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata.** *Nucleic Acids Res.* 2024 Jan 5; **52**(D1): D164–D173.
PubMed Abstract | Publisher Full Text | Free Full Text

76. 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

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

78. 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

79. 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

80. 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

81. Rasmussen JA, Chua PYS: **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; **271**: 127372.
Publisher Full Text

82. 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

83. 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

84. 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

85. 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

Open Peer Review

Current Peer Review Status: ✔ ✔ ?

Version 2

Reviewer Report 06 October 2025

https://doi.org/10.5256/f1000research.185171.r412934

?

Lauren Lui
E O Lawrence Berkeley National Laboratory, Berkeley, California, USA

Summary
Finn et al. have written a comprehensive white paper on the establishment of the ELIXIR Microbiome Community. They outline the purpose of establishing the ELIXIR Microbiome Community, the history of its formation, interaction with other ELIXIR communities, plans for data management and computational resources, and challenges in implementation. I found the manuscript informative about the intent of the ELIXIR Microbiome Community initiatives, and it was useful to have the historical and international context included. There is a lot of information packed into this manuscript! I only have minor issues with the manuscript that I think can easily be addressed.

Major Comments
On page 6, the authors state that the ELIXIR Microbiome Community is about providing the necessary infrastructure for "nucleotide sequence data" from a microbiome. In the context of the entire manuscript this statement is confusing because proteomics and metabolomics are also discussed in relation to the entire ELIXIR metacommunity. Should this statement be revised or can there be some clarification added here?

Why I put "no" for "Are all factual statements correct and adequately supported by citations?" and why I put "no" for "Are arguments sufficiently supported by evidence from the published literature"

There are some blanket statements in the manuscript that are consistent with the literature but there aren't any citations. There are a few statements that are incorrect.

- Introduction. There are many statements here that could be supported by literature references. For example, the statement that viruses that infect bacteria play critical roles in community dynamics could easily have a few references. This statement could also be expanded to include archaeal and eukaryotic viruses. For example, viruses that infect eukaryotic algae are also very important in the dynamics of algal blooms. The statement about the impact of dysbiosis of microbiomes could also have a few references.

- One Health. One Health is used in the manuscript but not defined nor does it have a citation. Please describe briefly and include a citation. Page 8.
- Microbial dark matter. There's at least two places that mentions "microbial dark matter" but don't define it (pages 8, and 17). Capitalization is not consistent between these two instances. Please include a reference and/or define this term.
- Page 16. Most of the content on this page could use more references:
  - Section on long reads.
    - This secion could include citations on long read metagenomics studies, notably the publication of Microflora Danica from Mads Albertsen's lab (https://www.nature.com/articles/s41564-025-02062-z). I acknowledge that this came out fairly recently and the authors may have drafted this section before this publication came out, but there are many other long read metagenome papers that can also be cited.
    - My second issue with this section is that long reads don't really mitigate the computational burden – you still can't assemble the metagenomes on a laptop and nanopore basecalling needs GPUs. Long reads assist with genome assembly because the reads are longer than genomic repeats, eliminating the ambiguities of short read assembly. This is discussed by Ryan Wick on his github page (https://github.com/rrwick/Unicycler?tab=readme-ov-file) and in the introduction and conclusion of this paper https://pmc.ncbi.nlm.nih.gov/articles/PMC8172020/.
  - Paragraph on MAGs.
    - MAGs should be defined as "metagenome-assembled genomes" not "environmental genomes." MAGs can also be from clinical metagenomes.
    - In the same sentence as above, the statement "has allowed the identification of thousands of specific functions" is unclear as to what it refers to. If this refers to protein gene function, then all of these would be by computational prediction and not really identification of new functions. A citation would be helpful here.
  - Paragraph about classification of MAGs.
    - Please add "taxonomic" in front of "classification" in the first sentence for clarity
    - There's been huge changes in prokaryotic taxonomy in recent years, and there are no citations to the committee reports here or opinion pieces of the use of "candidatus" when naming organisms. It is worth a few sentences of discussion here or added references.
  - Paragraph on viruses
    - Please add a few references for the statements about the challenges of analyzing viral microbiomes.
    - The statement that there is no centralized database is incorrect. IMG/VR, which is mentioned in the article, has millions of viral genomes.

Some statements may need to be qualified or rephrased.

- Page 11, the statement "Even relatively simple workflows that perform metagenomics assembly are computationally heavy." This statement is heavily qualified based on someone's experience with and access to computational resources, as well as the amount of data to be processed. It would also be better to use something like "resource-intensive" instead of "heavy." If this statement could also be changed to something like "most metagenomes cannot be assembled on a laptop" to provide more context for the reader.
- Page 17, last paragraph. The sentence about the challenges in metaproteomics mentions

proteins that are not identified in corresponding metagenomes and calls them “so called unidentified proteins of unknown function.” This seems to be a bit of a non sequitur because proteins of unknown function are those that cannot have a function assigned computationally, not because they are not found in the genomic data. It is also possible that "so called unidentified proteins of unknown function" is a separate list item - if so then numbers should be added into this sentence for clarity.

- Page 14, the statement beginning with “For example, it is widely accepted that current short-read assembly-based methods…” The accuracy of this statement depends on access to computational resources. Terabase scale metagenomes have been done (https://www.nature.com/articles/s41598-020-67416-5). Perhaps change “computationally intractable” to “highly computationally resource intensive.”

Minor comments

Use of “ELIXIR Microbiome Community”

- In most places, “ELIXIR Microbiome Community” is used, but in some places it is only “Microbiome Community.” Please change all of these instances to “ELIXIR Microbiome Community” for consistency.

ELIXIR Platforms

- Please add a table of what each of the platforms do or describe briefly in the text (top of page 5)

ASV

- ASV is the abbreviation for “amplicon sequence variant” not “amplified sequence variant”

Typos

- Page 11, “non-sequenced” should be “non-sequence”’
- Page 11, bottom of paragraph 4, should microbiome be lower case?

Citations

- INSDC needs a citation (bottom of page 5)

Figure 1

- I assume that the starting point for this diagram is the “Microbiomes in their environment” box. This could either be a unique color (so something other than white, blue, or orange) or larger than the other boxes to visually indicate that this is the starting point of the data/material workflow.

Abbreviations

Given the nature of this article, there are many acronyms and the definitions of a few were missed.

- GBIF – page 8
- Country abbreviations in Table 3 (FR, IT, DE)

Grammar

There are some grammatical issues sprinkled throughout the manuscript but are easily fixed:

1. Colons and semicolons are misused throughout this manuscript given the large number of lists. Colons only come after complete clauses. Semi-colons can only separate two independent clauses. For example, on page 4, the last sentence of the first paragraph should be something like “Key challenges facing the microbiome research community are

how to (i) appropriately store the data, (ii) informatically process, integrate, compare and interpret microbiome-derived data, and (iii) how to make the data findable, accessible, interoperable, and reusable, *i.e.,* FAIR." The colon was removed and commas used to separate the list items. This is also an issue in the third paragraph of section 2.2.2.5 Training, first paragraph of section 3, page 16 fifth paragraph, fourth paragraph on page 17, last row of table 3.

2. Use commas after i.e. and e.g.
3. I think that last sentence of the third paragraph on page 6 may need to have an "and" in front of "to biotechnology".
4. Page 17, last paragraph. The sentence about the challenges in metaproteomics could use numbers for the items. As it is currently structured, the sentence is confusing to read.
5. Page 18. Second paragraph would benefit from the use of the oxford comma. Example is add a comma after "image data." The last sentence of this paragraph is awkwardly worded.

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
No

Are arguments sufficiently supported by evidence from the published literature?
No

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbial ecology, Bioinformatics, Metagenomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 16 September 2025

https://doi.org/10.5256/f1000research.185171.r412360

Charlie Pauvert
University Hospital of RWTH, Aachen, Germany

Thanks to the authors for addressing all my comments and suggestions. They successfully restructured and amended the tables 2 and 3 and the article for an increased impact to highlight how the scientific community can benefit from the ELIXIR Microbiome Community.

On a small minor note, it seems possible to acknowledge the authors of the MIA package more than the URL by citing their work see https://microbiome.github.io/mia/authors.html#citation (and the added citation #1 via the form).

References

1. Borman, T Ernst, F Shetty, S Lahti, et al.: mia: Microbiome analysis. R package version 1.17.9. *https://microbiome.github.io/mia/*. 2025.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: microbiome, bioinformatics, reproducible research, data and metadata standards

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 11 May 2024

https://doi.org/10.5256/f1000research.158321.r251099

Almut Heinken
[1] University of Lorraine, Lorraine, France
[2] University of Lorraine, Lorraine, France

In this article, the ELIXIR Microbiome community is described. The community was established in 2015 as the Marine Metagenomics community and was recently expanded to the scope of metagenomics research for multiple areas such as human health, agriculture, and ecology. The authors then describe their perspective for the community. One focus area will be the promotion of best practices for metagenomics analyses by benchmarking tools. Interoperability between different tools will also be facilitated. Another focus will be encouraging data sharing and reuse and providing data storage platforms. Finally, links to existing ELIXIR initiatives in related areas as well as with international initiatives will be established.
Overall, this is a very clear, concise, and informative review. The scope and aims of ELIXIR Microbiome are well-described and detailed. The short-term and long-term objectives are also

clearly described.

Specific comments:

- I appreciate the links to other ELIXIR initiatives in Table 2. I would be particularly interested in more detail on the integration with the Systems Biology community. How would the ability to reuse multi-omics data for systems biology approaches be improved?
- It is a bit unclear to me which is the proposed data platforms, tools, and training will be freely available to non-ELIXIR members.
- Regarding providing metadata of samples, how will GDPR regulations be handled for human samples?

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Are arguments sufficiently supported by evidence from the published literature?

Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Systems biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 15 Jul 2025

Bérénice Batut

In this article, the ELIXIR Microbiome community is described. The community was established in 2015 as the Marine Metagenomics community and was recently expanded to the scope of metagenomics research for multiple areas such as human health, agriculture, and ecology. The authors then describe their perspective for the community. One focus area will be the promotion of best practices for metagenomics analyses by benchmarking tools. Interoperability between different tools will also be facilitated. Another focus will be encouraging data sharing and reuse and providing data storage platforms. Finally, links to existing ELIXIR initiatives in related areas as well as with international initiatives will be established.
Overall, this is a very clear, concise, and informative review. The scope and aims of ELIXIR Microbiome are well-described and detailed. The short-term and long-term objectives are also clearly described.

We would like to thank the reviewer for their positive comments concerning the ELIXIR microbiome community papers. Below we address their specific comments.

Specific comments:

- *I appreciate the links to other ELIXIR initiatives in Table 2. I would be particularly interested in more detail on the integration with the Systems Biology community. How would the ability to reuse multi-omics data for systems biology approaches be improved?*

We have added a few sentences to expand how the integration might be improved.

- *It is a bit unclear to me which is the proposed data platforms, tools, and training will be freely available to non-ELIXIR members.*

All of the proposed activities are freely available to non-ELIXIR members. ELIXIR does not put explicit boundaries on who can use, but engagement with some ELIXIR events may be preferentially given to scientists coming from ELIXIR member states and funds from ELIXIR funding schemes would be restricted to member states.

- *Regarding providing metadata of samples, how will GDPR regulations be handled for human samples?*

The landscape concerning human microbiome samples is complicated. Currently there is little consensus across Europe whether human microbiomes should be under controlled access. Similarly, the GDPR landscape is also complicated and not entirely independent. There are already established routes for suppression of data (should an individual wish to be forgotten), which can be propagated to other databases (e.g. MGnify will remove analyses associated with a suppressed sequence dataset). This is clearly an area that will need to be developed as part of the ELIXR community, as no specific solutions have been agreed, we would prefer not to comment about how they will be handled in this manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Report 14 March 2024

https://doi.org/10.5256/f1000research.158321.r244855

?

Charlie Pauvert

[1] University Hospital of RWTH, Aachen, Germany

[2] University Hospital of RWTH, Aachen, Germany

The authors make the case for mutating the ELIXIR Marine Metagenomics Community initiative into an ELIXIR Microbiome Community. It is indeed timely to have an such a pan-European initiative especially as it draws on previous experience, albeit a more narrow biome. The emphasis on multi-omics integration is also a strong suit as it is a complex topic where the microbiome research community would need support and infrastructure. The authors map out existing (but

not all) similar initiatives, even if it is unclear how they would work together. In my opinion, a couple of claims should be strengthened and the argumentative power of this white paper would benefit from minor reorganization, a text slim down and extra proofreading. To this end, I highlight the strengths and weaknesses of the manuscript below.

Strengths

- Table 3 is a great idea to map out the others initiatives but needs a bit rework to be straight to the point.
- In the Data section, the part between "Throughout the lifetime of the ELIXIR Marine Metagenomics Community" and the end of the paragraph is really relevant and draws from the experience of the ELIXIR Community. These sentences should be more highlighted, maybe upstream. I do wonder how the authors considered how the others initiatives (mentioned in Table 3) could also contribute to the addition of metadata for a global curation effort, possibly via the mentioned Contextual Data Clearinghouse.
- The data re-analysis and integration between PRIDE and MGnify (named MetaPUF) is a good use case of the efforts that the new ELIXIR Community can promote.
- I did appreciate the clear objective at the end of the Compute section, to help with scaling up analyses as well as the plan for the interaction with the ELIXIR Training Platform which promise to give a boost in training the current and next generation of microbiome researchers.
- I liked how the authors highlighted (using the example of viral sequences databases) that time and resources should not be wasted in duplicating works, and that initiatives such as the ELIXIR Microbiome Community can promote this.
- The authors thought ahead to promote and develop methods to limit confounding factors when re-using datasets, especially in multi-omics settings, and added this to one of the ELIXIR Microbiome Community challenges.

Weaknesses

Introduction

- (minor) The microbiome definition used in the paper was revisited in Review reference 1 to include the interactions as well as others molecules than genomes as part of the microbiome. I suggest to use this definition given the emphasis on additional omics made in the paper.
- (major) The first paragraph ends on challenges of our research community including how to make data FAIR. Given the experience gained by the ELIXIR Marine Metagenomics Community, I would have appreciated a sentence on why making our data FAIR is important (e.g., transparency to make the research more reproducible, accountability given the (public) source of funding). See major comment in next section.

The scope of the ELIXIR Microbiome Community

- (major) These following arguments for FAIR data should be in the Introduction section. "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."
- (minor) The very first sentence of this section blames researchers for misuse of a term "[...] regularly (incorrectly) used [...]". I suggest to rephrase the sentence to simply state the differences, or back up the misuse of the term with references or a survey.
- (minor) "Fundamentally, the ELIXIR Microbiome Community is about providing the necessary infrastructures required to perform analysis of **nucleotide** sequence data derived from a microbiome"
- (minor) "Finally, and possible unique to this ELIXIR Community, is the variety of researchers". I am not sure that these features are unique to the ELIXIR Community but rather to the field of microbiome research. I suggest to tone it down by simply pointing out that the ELIXIR Community gathers a variety of researchers.

Table 1

- (major) Metabolomics is listed in Table 1 but is omitted from the narrative in the paragraph.
- (major) The Table 1 is not really used but could actually reduce some redundancies in the manuscript by providing a one-stop-shop to explain and detail these techniques.

Figure 1

- (major) The figure is early on in the manuscript and it is unclear which of the ELIXIR platforms and communities are already established, or foreseen. Especially since "current and future ELIXIR activities" is mentioned in the manuscript before referencing this figure. The Table 2 does not add more information in that regard.
- (major) There is a need for different types of arrows as the same arrow represent interactions between communities or processes.
- (minor) The "Data" node is quite generic, I was wondering whether it meant public repositories, institute repositories or both. Please be precise.
- (minor) I guess the type of metadata illustrated in the figure is restricted to biological metadata, that is indeed collected during sampling, however, technical metadata such as the sequencing method used, the type of instrument or the library layout, are collected during the *processing* of the sample not only the sampling itself.
- (minor) There should be an arrow from "Archive" back to "Data" when the data produced is deposited and then contributes back to public repositories?

Interactions with other ELIXIR Communities

* (major) The sentence "Similarly, many of the biodiversity approaches use marker gene amplification for studying environmental DNA (eDNA)" is redundant with the previous one **and** introduces a different nomenclature that was not used before (marker gene amplication vs metabarcoding) and the differences, if any, are not explained. I would suggest to remove the sentence.
* (major) The term "Isolation of genomes" is misleading and I guess the authors used it as a shorthand for "The isolation of bacteria, its DNA extraction, genome sequencing and their annotation". Please rephrase to avoid misinterpretation. Plus I would argue that these steps are also done when deconstructing microbiomes via cultivation strategies and are therefore not so out-of-scope.

* (minor) "yet each one of these areas is far greater in scientific scope" feels exaggerated and vague. It should be rephrased. A suggestion is: "yet each one of these areas is too complex to be tacked individually"
* (minor) There is no link nor transition between the paragraph that starts with "In summary, " before the Table 3 and the paragraph after that starts with "Similarly, microbiome research". Please rephrase or edit to connect the two sections. Plus, whilst this is good to have a concrete example in the "Similarly" paragraph, the paragraph before was very broad and doing a summary. I would suggest to try reordering the two paragraphs and bring the example earlier for a smoother transition.
* (minor) Typo "Similarly, microbiome research has many translation**al** aspects"
* (minor) The PET acronym is detailed but actually used only once, I am unsure if this is necessary.

Table 2

- (major) The text in the "Interaction" column needs rework as it does not use a consistent wording and could be more to the point, especially as it is a complement to the main text:
  - The Food and Nutrition entry is a question.
  - The Galaxy entry has an unnecessary return carriage, and a unspecific "ongoing evaluation study" that needs to be clarified.
  - The Plant Science entry starts with a generic sentence that could be in the introduction or removed for clarity. Plus I would add a clarification that "plants maintain **or not** their microbial communities across generations."
- (minor) The status (e.g., currently active, inactive, planned, etc.) of each ELIXIR Communities would have been appreciated as it is missing also from the Figure 1.
- (minor) The mention of "the field" for the Federated human data community is vague in a manuscript about gathering communities, which research field is implied? If this is the human microbiome research field as a whole, please indicate.

Table 3

- (major) Similarly to the Table 2 major comment, there is a lack of consistent wording in the "Aim" column that makes the Table 3 not as impactful as it should and could be. Maybe the authors could extract common/distinct features from each of these initiatives as an alternative way to the "Aim" column. A couple of suggestions for these features would be: is it a national initiative?, does it relates to data storage, data analysis, training? is it linked to ELIXIR?
- (major) I was surprised not to see the NMDC listed in the table 3, especially when it is discussed in the main text. What about the NCCR Microbiomes initiative in Switzerland? I can understand that some initiatives are not included for space reason, but maybe state it in the legend of the table.
- (minor) Is this table sorted? It seems not, but it could be by acronyms or names.
- (minor) The aim for the NFDI4Microbiota is way too big a paragraph. The authors should reduce it for conciseness.
- (minor) The Metaproteomics Initiative entry has a hyperlink and a reference when none of the others have. Please homogenise.
- (minor) Some entries have country listed and some not. Please homogenise.
- (minor) I am not questioning the existence of the European Reference Genome Atlas here, but how best to phrase its relevance to ELIXIR in the manuscript. There seems to have no prokaryotes genomes in their atlas, however, there seems to be a trove of fungi and protists genomes which

are usually said to be understudied in microbiome. So I think there is a missed opportunity for the authors here to make the most out of this entry.
- (minor) "With its headquarters in Bari (Apulia region)," seems irrelevant in the context of the table, please remove.

Data

- (minor) The end of the following sentence is redundant as the INSDC was introduced earlier already "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.". Please adjust accordingly.
- (minor) "A current challenge facing the field is connecting different multi 'omics data that have been derived from the same sample." Is this going to be tackled by the ELIXIR Microbiome Community? If so, I would state that this is part of its objectives.
- (minor) The end of sentence "[...]overarching context to the experiment, which can be important for meta-analyses." seems like an euphemism, I would suggest to replace with "[...]overarching context to the experiment, re-analyses or meta-analyses." to include re-analyses as well.
- (minor) "We will continue to promote such approaches, enriching metadata wherever possible." Is this going to be done via the CDCH?
- (minor) "The ELIXIR Microbiome Community will also work to move the Marine Metagenomics domain in the RDMKit towards a more general Microbiome domain." What is the RDMKit? It is not explained, nor cited nor mentioned again.

Tools

- (minor) "will increase their use of BioContainers" should be "will increase the use of BioContainers"
- (minor) "In order to make tools findable **by the end users,** the Community"
- (minor) "workflow descriptions (e.g. Snakemake, CWL, Nextflow)" None of them have their references cited, is it an omission or space limitation?
- (minor) "A **current** joint effort between the Microbiome and Galaxy Communities"

Benchmarking

- (major) "Benchmarking" it is the only item at this hierarchy level, meaning that this is useless for structuring the text. Please edit.
- (minor) Review reference 2 published a recent review with guidelines to learn from that could have its place in this paragraph.
- (minor) in the sentence: "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.", is this the ELIXIR Microbiome Community, or the wide community of microbiome researchers? Is it to mean that the ELIXIR Microbiome Community is going to act as an interface between microbiome researchers and ELIXIR Tools/Infrastructure?

Compute
- (minor) The first sentence would fit better in the introduction.

Interoperability
- (minor) The reference 57 should be at the end of the sentence starting with "This effort was paralleled" not in the middle.
- (minor) Reference 58 should be removed at it is a duplicate of reference 47.
- (minor) Use the full text "Global Alliance for Genomics and Health" instead of GA4GH.
- (minor) "is in the process of applying to be a**n ELIXIR** Recommended Interoperability Resource."
- (minor) In the sentence: "This will require the development of new data Interoperability layers for data resources that are not normally focused in Microbiome data" I think the authors meant "used" instead of "focused", and "microbiome" instead of "Microbiome".

Training
- (minor) In "Platforms such as MGnify support large-scale services for most, if not all, steps of a microbiome study", I would suggest to remove the "if not all".
- (minor) "with areas of expertise covering different environments, ‘omics approaches and data analysis pathways." I think the authors meant "learning paths", and I would refrain from using "pathways" as it also has a biological meaning.

Context with other international initiatives

- (major) Whilst I appreciated the emphasis that no initiative exists on its own, I feel the first paragraph on the Genom**ic** (please correct the typo) Standards Consortium feels lengthy for a manuscript whose topic is not the GSC. I would advise to summarize. In this respect, the second paragraph is particularly relevant to a tangible collaboration between ELIXIR Microbiome Community and GSC.
- (major) The NMDC is discussed in this section but not part of the Table 3.
- (minor) Is the mentioned M5 project still active as the website's last update is 2012?
- (minor) "Combining the activities on standards concerning workflows [...]" does this means adding and providing workflows to the microbiome research community?
- (minor) Given the emphasis on the fact that MicrobiomeSupport *was* a program, the authors could update the readers and indicate that it is now MicrobiomeSupport Association.

Interaction with other key data resources beyond ELIXIR

- (major) There is an order issue with the main text that a proofread could solve, as MG-RAST is explained and cited in the first paragraph but already mentioned upstream of the main text in the Interoperability section.

Specific challenges and objectives of the ELIXIR Microbiome Community

- (major) The strong claim "it is *widely accepted* that current short-read assembly-based methods do not generally work as well for soil microbiomes" would probably need at least one reference.
- (major) "(iii) there is no centralised database collecting the millions of viral sequences". It seems to be the case indeed, and there are databases (~24) out there as recently compiled in Review reference 3. How ELIXIR Microbiome Community plans to integrate/aggregate these resources in a non-duplicating manner?
- (minor) The word "through" is superfluous in the the sentence that starts with "This current limitation, [...]" and can be removed.

- (minor) The CAMI was already explained and cited above, so the already defined acronym can be used.
- (minor) Precise the area in :"Additionally, another key area of development **of taxonomy** [...]".
- (minor) The sentence "Viruses, particularly those that infect bacteria, are found ubiquitously in all environments and play critical roles in community dynamics." belongs in an introduction, not so downstream of the manuscript.
- (minor) The sentence starting the sixth paragraph could be precised as "The increase in metagenomic assemblies has resulted in a parallel increase in the number of **predicted** protein sequences, with sets of non-redundant proteins now in the billions."
- (minor) Fix typo in "that are undetectable by current sequence based methods."
- (minor) Would it make sense to also be able to access representative, of clusters for instance? "[...] develop new infrastructural frameworks for accessing slices of the data **or adequate representatives** based on the requirements."
- (minor) What is the "**expanded** Microbiome Community"? It was never mentioned before.
- (minor) A few comments on the sentence: "Within the Community, we will develop and promote standards around the analysis provenance (analytical metadata),". In my opinion and how it was already stated in the manuscript, it would make more sense to promote existing standards first and then develop if need be. There is no mention of other type of metadata, so the "analytical metadata" precision seems superfluous. I would suggest: "Within the Community, we will **promote and develop** standards **regarding** the analysis provenance,"
- (minor) Precise the term forms in "[...] ensuring that comput**ing** resources are accessible for performing the different *forms* of data analysis [...]", do the authors mean types of/steps in the data analysis?
- (minor) Same argument as before regarding reinventing the wheel, I would swap the part of the sentence: "This may require the **extensions to existing databases or development of new ones**, but it requires an agreement from the research community to adopt them."
- (minor) This part "Metaproteomics aims to elucidate the functional and taxonomic interplay of proteins in microbiomes," should have been in the Table 1, or to reuse the Table 1 here.
- (minor) The tenth paragraph of this section starts with the mention of multiple major challenges, but detail "only" one of them. I would suggest to mention some of the others challenges.
- (minor) The "MIA" method is just a hyperlink, without any reference. Either cite the website accordingly or add the reference.

Table 4

- (major) I was surprised to see that the objective "Foster international collaborations between other resources providers and databases to ensure global harmonisation of e-infrastructures for microbiome research" was long-term, as I would have imagined that a gap analysis would be short-term to ensure we do not reinvent the wheel, especially given the others initiatives discussed in the manuscript.
- (minor) If the Objective column starts with action verbs (which is a good idea), then it should be "Survey the needs" instead of "Survey of needs".
- (minor) Specify the type of workflow with "Address knowledge gaps in generating and adopting **data analysis** workflows"
- (minor) The verb is missing in "**Teach a**dvanced containerisation and cloud deployment"
- (minor) The verb is missing in "**Promote** data analysis through the use of services". Is this ELIXIR services in general or specific ELIXIR Microbiome Community services?
- (minor) Use "Share" instead of "Sharing" in the "Co-ordinate" entry.

- (minor) The "Industry connection" entry does not fit the action verb pattern. A suggestion would be "Use ELIXIR and Node forums to understand pharmaceutical and biotechnological demands and current limitations impacting this sector." as the first sentence felt generic.
- (minor) The verb is missing in "**Design** targeted training for different microbiome communities"
- (minor) Use "findability" instead of "discoverability" for consistency and to fit with the FAIR principles.
- (minor) Reorder the sentence to start with the action verb: "**E**stablish new standards for microbiome research, particularly with respect to data analysis reporting and contextual metadata reporting **in conjunction with GSC**"
- (minor) Correct "Established" to "Establish" in the "Promoting new approaches" entry.

Editorial comments:
- (major) "Community" is used in upper-case and this is unclear in many instances whether the ELIXIR Marine Metagenomics Community is referred to, the ELIXIR Microbiome Community, or the broader microbiome research community. I suggest to use consistently the full term for the sake of transparency. Abbreviations like EMMC and EMC could be even more misleading in my opinion.
- (major) The structure of the white paper is not evident as the hierarchy is indicated only by change in font size, and some sections are quite lengthy for a white paper that is supposed to be concise. I understand that this is a constraint from the Editor, but see Review reference 4 for a white paper with a more clearer structure. An alternative could be to use numbered sections.

- References
  - (minor) The very first reference is oddly formatted in the text creating an artificial and confusing end of sentence.
  - (minor) Title of reference 9 is truncated and should be "Methods included: standardizing computational reuse and portability with the Common Workflow Language"
  - (minor) The superscript numbers of the numeric style of bibliography are in many instances after the final dot (see reference 2, 12-16, 17-19, 36, 37), after a comma (see reference 10, 20, 23), a bracket (see reference 60) or semi colon (see reference 65) when they should be before any of these symbols.
  - (minor) In the paragraph "Interactions with other ELIXIR Communities", we jump from reference 24 to 29 when the numeric style of bibliography (that was chosen by the authors) is expected to mirror the mentions in the manuscript. Please either have the Table 2 earlier in the paper, or change the order of the references.

References

1. Berg G, Rybakova D, Fischer D, Cernava T, et al.: Microbiome definition re-visited: old concepts and new challenges. *Microbiome*. 2020; **8** (1). Publisher Full Text
2. Ritsch M, Cassman NA, Saghaei S, Marz M: Navigating the Landscape: A Comprehensive Review of Current Virus Databases.*Viruses*. 2023; **15** (9). PubMed Abstract | Publisher Full Text
3. Brooks TG, Lahens NF, Mrčela A, Grant GR: Challenges and best practices in omics benchmarking.*Nat Rev Genet*. 2024. PubMed Abstract | Publisher Full Text
4. Vizcaíno JA, Walzer M, Jiménez RC, Bittremieux W, et al.: A community proposal to integrate proteomics activities in ELIXIR.*F1000Res*. 2017; **6**. PubMed Abstract | Publisher Full Text

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?
Partly

Are arguments sufficiently supported by evidence from the published literature?
Partly

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: microbiome, bioinformatics, reproducible research, data and metadata standards

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 Jul 2025

Bérénice Batut

The authors make the case for mutating the ELIXIR Marine Metagenomics Community initiative into an ELIXIR Microbiome Community. It is indeed timely to have an such a pan-European initiative especially as it draws on previous experience, albeit a more narrow biome. The emphasis on multi-omics integration is also a strong suit as it is a complex topic where the microbiome research community would need support and infrastructure. The authors map out existing (but not all) similar initiatives, even if it is unclear how they would work together. In my opinion, a couple of claims should be strengthened and the argumentative power of this white paper would benefit from minor reorganization, a text slim down and extra proofreading. To this end, I highlight the strengths and weaknesses of the manuscript below.

Strengths

- Table 3 is a great idea to map out the others initiatives but needs a bit rework to be straight to the point.
- In the Data section, the part between "Throughout the lifetime of the ELIXIR Marine Metagenomics Community" and the end of the paragraph is really relevant and draws from the experience of the ELIXIR Community. These sentences should be more highlighted, maybe upstream. I do wonder how the authors considered how the others initiatives (mentioned in Table 3) could also contribute to the addition of metadata for a global curation effort, possibly via the mentioned Contextual Data Clearinghouse.
- The data re-analysis and integration between PRIDE and MGnify (named MetaPUF) is a good use case of the efforts that the new ELIXIR Community can promote.
- I did appreciate the clear objective at the end of the Compute section, to help with scaling up

analyses as well as the plan for the interaction with the ELIXIR Training Platform which promise to give a boost in training the current and next generation of microbiome researchers.
- I liked how the authors highlighted (using the example of viral sequences databases) that time and resources should not be wasted in duplicating works, and that initiatives such as the ELIXIR Microbiome Community can promote this.
- The authors thought ahead to promote and develop methods to limit confounding factors when re-using datasets, especially in multi-omics settings, and added this to one of the ELIXIR Microbiome Community challenges.

Weaknesses

Introduction

- (minor) The microbiome definition used in the paper was revisited in Review reference 1 to include the interactions as well as others molecules than genomes as part of the microbiome. I suggest to use this definition given the emphasis on additional omics made in the paper.

We have added that the definition includes other molecules in addition to genomes.

- (major) The first paragraph ends on challenges of our research community including how to make data FAIR. Given the experience gained by the ELIXIR Marine Metagenomics Community, I would have appreciated a sentence on why making our data FAIR is important (e.g., transparency to make the research more reproducible, accountability given the (public) source of funding). See major comment in next section.
Thank you for this suggestion. We have added a few sentences to this effect in the manuscript.

The scope of the ELIXIR Microbiome Community

- (major) These following arguments for FAIR data should be in the Introduction section. "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."

We have merged this part of the paragraph into the Introduction.

- (minor) The very first sentence of this section blames researchers for misuse of a term "[...] regularly (incorrectly) used [...]". I suggest to rephrase the sentence to simply state the differences, or back up the misuse of the term with references or a survey.

We have qualified this with an example of INSDC mislabelling.

- (minor) "Fundamentally, the ELIXIR Microbiome Community is about providing the necessary

*infrastructures required to perform analysis of **nucleotide** sequence data derived from a microbiome"*

“Nucleotide” has been added to this sentence.

- (minor) "Finally, and possible unique to this ELIXIR Community, is the variety of researchers". I am not sure that these features are unique to the ELIXIR Community but rather to the field of microbiome research. I suggest to tone it down by simply pointing out that the ELIXIR Community gathers a variety of researchers.

Modified accordingly.

Table 1

- (major) Metabolomics is listed in Table 1 but is omitted from the narrative in the paragraph.

We have added a sentence in the paragraph.

- (major) The Table 1 is not really used but could actually reduce some redundancies in the manuscript by providing a one-stop-shop to explain and detail these techniques.

We have increased the cross linking to the table.

Figure 1

- (major) The figure is early on in the manuscript and it is unclear which of the ELIXIR platforms and communities are already established, or foreseen. Especially since "current and future ELIXIR activities" is mentioned in the manuscript before referencing this figure. The Table 2 does not add more information in that regard.

- (major) There is a need for different types of arrows as the same arrow represent interactions between communities or processes.

Modified accordingly.

- (minor) The "Data" node is quite generic, I was wondering whether it meant public repositories, institute repositories or both. Please be precise.

"Data" refers here to the ELIXIR Data Platform

- (minor) I guess the type of metadata illustrated in the figure is restricted to biological metadata, that is indeed collected during sampling, however, technical metadata such as the sequencing method used, the type of instrument or the library layout, are collected during the processing of the sample not only the sampling itself.

This is everything from phenotypic, sample conditions, experimental methods

- (minor) There should be an arrow from "Archive" back to "Data" when the data produced is deposited and then contributes back to public repositories?

Interactions with other ELIXIR Communities

** (major) The sentence "Similarly, many of the biodiversity approaches use marker gene amplification for studying environmental DNA (eDNA)" is redundant with the previous one* ***and*** *introduces a different nomenclature that was not used before (marker gene amplication vs metabarcoding) and the differences, if any, are not explained. I would suggest to remove the sentence.*

We have modified the sentence to refer to barcoding rather than removing the sentence. We feel it is important to mention eDNA.

** (major) The term "Isolation of genomes" is misleading and I guess the authors used it as a shorthand for "The isolation of bacteria, its DNA extraction, genome sequencing and their annotation". Please rephrase to avoid misinterpretation. Plus I would argue that these steps are also done when deconstructing microbiomes via cultivation strategies and are therefore not so out-of-scope.*

This sentence has been removed as the reviewer is correct that cultivation of bacteria (and other organisms) from microbiomes is becoming an increasing trend.

** (minor) "yet each one of these areas is far greater in scientific scope" feels exaggerated and vague. It should be rephrased. A suggestion is: "yet each one of these areas is too complex to be tacked individually"*

We have amended the sentence accordingly.

** (minor) There is no link nor transition between the paragraph that starts with "In summary, " before the Table 3 and the paragraph after that starts with "Similarly, microbiome research". Please rephrase or edit to connect the two sections. Plus, whilst this is good to have a concrete example in the "Similarly" paragraph, the paragraph before was very broad and doing a summary. I would suggest to try reordering the two paragraphs and bring the example earlier for a smoother transition.*

The paragraphs have been reordered as suggested and slightly amended to improve readability.

** (minor) Typo "Similarly, microbiome research has many translation****al*** *aspects"*

Fixed

** (minor) The PET acronym is detailed but actually used only once, I am unsure if this is necessary.*

There is another instance of this abbreviation, so we have retained this abbreviation.

Table 2

- (major) The text in the "Interaction" column needs rework as it does not use a consistent wording and could be more to the point, especially as it is a complement to the main text:
- The Food and Nutrition entry is a question.
- The Galaxy entry has an unnecessary return carriage, and a unspecific "ongoing evaluation study" that needs to be clarified.
*- The Plant Science entry starts with a generic sentence that could be in the introduction or removed for clarity. Plus I would add a clarification that "plants maintain **or not** their microbial communities across generations."*
- (minor) The status (e.g., currently active, inactive, planned, etc.) of each ELIXIR Communities would have been appreciated as it is missing also from the Figure 1.
- (minor) The mention of "the field" for the Federated human data community is vague in a manuscript about gathering communities, which research field is implied? If this is the human microbiome research field as a whole, please indicate.

In response to all of the above comments, we have substantially reworked table 2. The only comment that we have not addressed so specifically, is whether an activity is in progress or not, because the activities ebb and flow, and it is not always easy to say when there is a start or an end. However, being somewhat more focused in the content, we feel this revised table provides a stronger view of the direction that the community wants to take with the other communities.

Table 3

- (major) Similarly to the Table 2 major comment, there is a lack of consistent wording in the "Aim" column that makes the Table 3 not as impactful as it should and could be. Maybe the authors could extract common/distinct features from each of these initiatives as an alternative way to the "Aim" column. A couple of suggestions for these features would be: is it a national initiative?, does it relates to data storage, data analysis, training? is it linked to ELIXIR?
- (major) I was surprised not to see the NMDC listed in the table 3, especially when it is discussed in the main text. What about the NCCR Microbiomes initiative in Switzerland? I can understand that some initiatives are not included for space reason, but maybe state it in the legend of the table.
- (minor) Is this table sorted? It seems not, but it could be by acronyms or names.
- (minor) The aim for the NFDI4Microbiota is way too big a paragraph. The authors should reduce it for conciseness.
- (minor) The Metaproteomics Initiative entry has a hyperlink and a reference when none of the others have. Please homogenise.
- (minor) Some entries have country listed and some not. Please homogenise.
- (minor) I am not questioning the existence of the European Reference Genome Atlas here, but how best to phrase its relevance to ELIXIR in the manuscript. There seems to have no prokaryotes genomes in their atlas, however, there seems to be a trove of fungi and protists genomes which are usually said to be understudied in microbiome. So I think there is a missed opportunity for the authors here to make the most out of this entry.
- (minor) "With its headquarters in Bari (Apulia region)," seems irrelevant in the context of the

table, please remove.

As with Table 2, we have substantially reworked Table 3, including ordering by the reach of the effort, harmonising the language and reducing text. We have also included the relevance of the activity to the community. We have not included NMDC, as this is a US effort, and while important to the community, this table is focused on Europe efforts.

Data

- (minor) The end of the following sentence is redundant as the INSDC was introduced earlier already "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.". Please adjust accordingly.

We have removed the redundancy here.

- (minor) "A current challenge facing the field is connecting different multi 'omics data that have been derived from the same sample." Is this going to be tackled by the ELIXIR Microbiome Community? If so, I would state that this is part of its objectives.

We feel this is likely to be a joint effort across different communities. Thus, we have retained the sentence as is.

- (minor) The end of sentence "[...]overarching context to the experiment, which can be important for meta-analyses." seems like an euphemism, I would suggest to replace with "[...]overarching context to the experiment, re-analyses or meta-analyses." to include re-analyses as well.

We have added "re-analyses or" as suggested.

- (minor) "We will continue to promote such approaches, enriching metadata wherever possible." Is this going to be done via the CDCH?

While CDCH offers one approach, we believe that there are many different ways that metadata may be enriched, first by making scientists more aware of the need of submitting meta, promoting different, more specific checklists (e.g. STORMS or MicroB3) and mining metadata from published literature. This, specifically highlighting CDCH would not be appropriate.

- (minor) "The ELIXIR Microbiome Community will also work to move the Marine Metagenomics domain in the RDMKit towards a more general Microbiome domain." What is the RDMKit? It is not explained, nor cited nor mentioned again.

This has been expanded to the ELIXIR Research Data Management Kit and a link has been added to the text.

Tools

- (minor) "will increase their use of BioContainers" should be "will increase the use of BioContainers"

Corrected.

*- (minor) "In order to make tools findable **by the end users,** the Community"*

Done.

- (minor) "workflow descriptions (e.g. Snakemake, CWL, Nextflow)" None of them have their references cited, is it an omission or space limitation?

Added references

*- (minor) "A **current** joint effort between the Microbiome and Galaxy Communities"*

Fixed

Benchmarking

- (major) "Benchmarking" it is the only item at this hierarchy level, meaning that this is useless for structuring the text. Please edit.

Thank you for pointing this out. We have changed the sub-sub-heading into an introductory sentence to this paragraph.

- (minor) Review reference 2 published a recent review with guidelines to learn from that could have its place in this paragraph.

*- (minor) in the sentence: "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.", is this the ELIXIR Microbiome Community, or the wide community of microbiome researchers? Is it to mean that the ELIXIR Microbiome Community is going to act as an interface between microbiome researchers and ELIXIR Tools/Infrastructure?*

We clarified this to mean the wider community of microbiome researchers.

Compute

- (minor) The first sentence would fit better in the introduction.

Interoperability

- (minor) The reference 57 should be at the end of the sentence starting with "This effort was

paralleled" not in the middle.

This citation has been moved to the end of the sentence.

- (minor) Reference 58 should be removed at it is a duplicate of reference 47.

References have been fixed

- (minor) Use the full text "Global Alliance for Genomics and Health" instead of GA4GH.

This abbreviation has been expanded.

*- (minor) "is in the process of applying to be a**n ELIXIR** Recommended Interoperability Resource."*

ELIXIR has been added to this sentence.

- (minor) In the sentence: "This will require the development of new data Interoperability layers for data resources that are not normally focused in Microbiome data" I think the authors meant "used" instead of "focused", and "microbiome" instead of "Microbiome".

We have updated the sentence accordingly.

Training

- (minor) In "Platforms such as MGnify support large-scale services for most, if not all, steps of a microbiome study", I would suggest to remove the "if not all".

Agreed.

- (minor) "with areas of expertise covering different environments, 'omics approaches and data analysis pathways." I think the authors meant "learning paths", and I would refrain from using "pathways" as it also has a biological meaning.

Actually, we do mean data analysis, but have changed to data analysis strategies. This is in reference to the fact that there may be multiple different ways of analysing a data type, for example metagenomics can be analysed using tools such as Kraken or MetaPhlAn4, to full assembly, gene calling and functional analysis.

Context with other international initiatives

*- (major) Whilst I appreciated the emphasis that no initiative exists on its own, I feel the first paragraph on the Genom**ic** (please correct the typo) Standards Consortium feels lengthy for a manuscript whose topic is not the GSC. I would advise to summarize. In this respect, the second paragraph is particularly relevant to a tangible collaboration between ELIXIR Microbiome Community and GSC.*

The typo has been corrected.

- (major) The NMDC is discussed in this section but not part of the Table 3.

Table 3 lists pan-European efforts, whereas the NMDC is a US specific initiative. Thus, we have not added this to the table.

- (minor) Is the mentioned M5 project still active as the website's last update is 2012?

While the website has not been updated, there are ongoing efforts to expand and revitalise this effort. RDF is a member of the GSC board and is promoting aspects of the M5 initiative. While the name may change, we feel this is useful to keep. However, as part of condensing the overall length of the manuscript, we have removed the reference to this initiative.

- (minor) "Combining the activities on standards concerning workflows [...]" does this means adding and providing workflows to the microbiome research community?

- (minor) Given the emphasis on the fact that MicrobiomeSupport was a program, the authors could update the readers and indicate that it is now MicrobiomeSupport Association.

Interaction with other key data resources beyond ELIXIR

- (major) There is an order issue with the main text that a proofread could solve, as MG-RAST is explained and cited in the first paragraph but already mentioned upstream of the main text in the Interoperability section.

This has now been fixed.

Specific challenges and objectives of the ELIXIR Microbiome Community

- (major) The strong claim "it is *widely accepted* that current short-read assembly-based methods do not generally work as well for soil microbiomes" would probably need at least one reference.

Added

- (major) "(iii) there is no centralised database collecting the millions of viral sequences". It seems to be the case indeed, and there are databases (~24) out there as recently compiled in Review reference 3. How
ELIXIR Microbiome Community plans to integrate/aggregate these resources in a non-duplicating manner?

- (minor) The word "through" is superfluous in the the sentence that starts with "This current limitation, [...]" and can be removed.

We have removed the sentence “through”.

- (minor) The CAMI was already explained and cited above, so the already defined acronym can be used.

This has been fixed.

*- (minor) Precise the area in :"Additionally, another key area of development **of taxonomy** [...]".*

Added

- (minor) The sentence "Viruses, particularly those that infect bacteria, are found ubiquitously in all environments and play critical roles in community dynamics." belongs in an introduction, not so downstream of the manuscript.

*- (minor) The sentence starting the sixth paragraph could be precised as "The increase in metagenomic assemblies has resulted in a parallel increase in the number of **predicted** protein sequences, with sets of non-redundant proteins now in the billions."*

Added "predicted" to the sentence.

- (minor) Fix typo in "that are undetectable by current sequence based methods."

Fixed - sequenced -> sequence

*- (minor) Would it make sense to also be able to access representative, of clusters for instance? "[...] develop new infrastructural frameworks for accessing slices of the data **or adequate representatives** based on the requirements."*

Added

*- (minor) What is the "**expanded** Microbiome Community"? It was never mentioned before.*

Removed "expanded" from the sentence.

*- (minor) A few comments on the sentence: "Within the Community, we will develop and promote standards around the analysis provenance (analytical metadata),". In my opinion and how it was already stated in the manuscript, it would make more sense to promote existing standards first and then develop if need be. There is no mention of other type of metadata, so the "analytical metadata" precision seems superfluous. I would suggest: "Within the Community, we will **promote and develop** standards **regarding** the analysis provenance,"*

We agree with the reviewer's comment and have re-ordered accordingly.

*- (minor) Precise the term forms in "[...] ensuring that comput**ing** resources are accessible for performing the different forms of data analysis [...]", do the authors mean types of/steps in the data analysis?*

"compute resources" is an accepted resource. We have removed the "different forms

of" as we feel it is a little superfluous. We were meaning the different strategies, but it does not add to the sentence.

- (minor) Same argument as before regarding reinventing the wheel, I would swap the part of the sentence:
*"This may require the **extensions to existing databases or development of new ones**, but it requires an agreement from the research community to adopt them."*

We agree, and have amended accordingly.

- (minor) This part "Metaproteomics aims to elucidate the functional and taxonomic interplay of proteins in microbiomes," should have been in the Table 1, or to reuse the Table 1 here.

We have highlighted this in table 1. We have kept the text here to provide the context for the rest of the sentence.

- (minor) The tenth paragraph of this section starts with the mention of multiple major challenges, but detail "only" one of them. I would suggest to mention some of the others challenges.

- (minor) The "MIA" method is just a hyperlink, without any reference. Either cite the website accordingly or add the reference.

There is not a reference, and the hyperlink goes to the GitHub site, as requested by the authors. We have included "https://github.com/microbiome/mia" for completeness sake.

Table 4

- (major) I was surprised to see that the objective "Foster international collaborations between other resources providers and databases to ensure global harmonisation of e-infrastructures for microbiome research" was long-term, as I would have imagined that a gap analysis would be short-term to ensure we do not reinvent the wheel, especially given the others initiatives discussed in the manuscript.

- (minor) If the Objective column starts with action verbs (which is a good idea), then it should be "Survey the needs" instead of "Survey of needs".

We have amended accordingly.

*- (minor) Specify the type of workflow with "Address knowledge gaps in generating and adopting **data analysis** workflows"*

Added

*- (minor) The verb is missing in "**Teach a**dvanced containerisation and cloud deployment"*

Added.

*- (minor) The verb is missing in "**Promote** data analysis through the use of services". Is this ELIXIR services in general or specific ELIXIR Microbiome Community services?*

Added Promote - generalised to ELIXIR services.

- (minor) Use "Share" instead of "Sharing" in the "Co-ordinate" entry.

Done

- (minor) The "Industry connection" entry does not fit the action verb pattern. A suggestion would be "Use ELIXIR and Node forums to understand pharmaceutical and biotechnological demands and current limitations impacting this sector." as the first sentence felt generic.

Done

*- (minor) The verb is missing in "**Design** targeted training for different microbiome communities"*

Done

- (minor) Use "findability" instead of "discoverability" for consistency and to fit with the FAIR principles.

Done.

*- (minor) Reorder the sentence to start with the action verb: "**E**stablish new standards for microbiome research, particularly with respect to data analysis reporting and contextual metadata reporting **in conjunction with GSC**"*

Done

- (minor) Correct "Established" to "Establish" in the "Promoting new approaches" entry.

Done

Editorial comments:

- (major) "Community" is used in upper-case and this is unclear in many instances whether the ELIXIR Marine Metagenomics Community is referred to, the ELIXIR Microbiome Community, or the broader microbiome research community. I suggest to use consistently the full term for the sake of transparency. Abbreviations like EMMC and EMC could be even more misleading in my opinion.

We have prefixed all instances of "Community" with their explicit community name.

- (major) The structure of the white paper is not evident as the hierarchy is indicated only by change in font size, and some sections are quite lengthy for a white paper that is supposed to be

concise. I understand that this is a constraint from the Editor, but see Review reference 4 for a white paper with a more clearer structure. An alternative could be to use numbered sections.

We have introduced section numbers to aid the structuring of the manuscript.

- References
- (minor) The very first reference is oddly formatted in the text creating an artificial and confusing end of sentence.
- (minor) Title of reference 9 is truncated and should be "Methods included: standardizing computational reuse and portability with the Common Workflow Language"
- (minor) The superscript numbers of the numeric style of bibliography are in many instances after the final dot (see reference 2, 12-16, 17-19, 36, 37), after a comma (see reference 10, 20, 23), a bracket (see reference 60) or semi colon (see reference 65) when they should be before any of these symbols.
- (minor) In the paragraph "Interactions with other ELIXIR Communities", we jump from reference 24 to 29 when the numeric style of bibliography (that was chosen by the authors) is expected to mirror the mentions in the manuscript. Please either have the Table 2 earlier in the paper, or change the order of the references.

References have been fixed

Competing Interests: No competing interests were disclosed.