MMSL
Military Medical Science Letters
Since 1925

Mil. Med. Sci. Lett. (Voj. Zdrav. Listy) 2026, 95(1), 33-46
ISSN 2571-113X (Online)
DOI: 10.31482/mmsl.2025.003

REVIEW ARTICLE

RESILIENT PLATFORM FOR MICROBIOME DATABASES: ESSENTIAL ATTRIBUTES

Tomáš Ráčil ✉, Petr Františ, Alexandr Štefek

Department of Informatics and Cyber Operations, University of Defence, Brno, Czech Republic

Received 20th April 2025.
Accepted 9th July 2025.
Published 2nd March 2026.

Summary

The rapid growth of microbiome research is hindered by significant challenges in data management, including data fragmentation across disparate silos, a lack of methodological standardization, and barriers to advanced privacy-preserving analysis. To address these issues, this article proposes a conceptual architectural blueprint for a resilient, scalable, and integrated platform for microbiome data. Our proposed architecture is a modular, cloud-native system designed to support the entire research lifecycle. Key attributes include a multi-layered microservices framework to ensure scalability, adherence to FAIR (Findable, Accessible, Interoperable, Reusable) data principles, and native support for longitudinal data tracking. Crucially, the platform incorporates integrated services for advanced AI/ML analysis and a coordinator for federated learning, enabling collaborative model development without centralizing sensitive data. By providing a robust infrastructure that combines standardized data management with powerful, privacy-aware analytical tools, our proposed model aims to empower researchers, enhance reproducibility, and accelerate discoveries into the complex relationship between the microbiome and human health.

Key words: Microbiome Database; GMrepo; Data Sharing in Microbiology; Bioinformatics Tools; Database Interoperability

1. Introduction

The human gut microbiome is more than a collection of resident microorganisms. This complex and dynamic ecosystem, teeming with trillions of microorganisms encompassing bacteria, archaea, fungi, and viruses, has emerged as a key player in influencing various aspects of human health (1, 2).

Estimated to harbor over one thousand bacterial species alone (3), the gut microbiome exhibits remarkable inter-individual variability. This remarkable inter-individual variability, stemming from factors such as diet, lifestyle, and genetics, poses a significant challenge to comprehensively understanding the gut microbiome's influence on human health. Researchers are continuously striving to decipher the intricate interplay between specific microbial communities and their functional roles, but the vast diversity complicates efforts to identify statistically significant patterns. Despite the power of machine learning in analyzing large datasets, its effectiveness diminishes when applied to small, intricate samples (5).

✉ University of Defence, Department of Informatics and Cyber Operations, Kounicova 65, 662 10 Brno, Czech Republic
tomas.racil@unob.cz

Despite advancements in sequencing technologies, studying the gut microbiome presents unique challenges. One major hurdle lies in the small-scale nature of microbiome samples, which often contain limited amounts of DNA, making them ill-suited for traditional analysis methods. Additionally, the vast diversity and inter-individual variability within the gut microbiome further complicates efforts to identify statistically significant patterns and draw definitive conclusions about the specific contributions of individual microbial species or communities to human health.

Furthermore, the application of machine learning to analyze vast microbiome datasets encounters limitations. While it has revolutionized data analysis in various fields, machine learning algorithms are susceptible to overfitting when dealing with small, intricate datasets like those obtained from the gut microbiome. Overfitting occurs when a model becomes overly adapted to the specific training data, hindering its ability to generalize effectively to new, unseen data, ultimately leading to unreliable predictions (6).

To surmount these challenges, estabilishing centralized, expansive data repositories for the gut microbiome holds immense promise (7). These repositories enable the pooling of data from various studies, facilitating the creation of larger, more comprehensive datasets. These comprehensive datasets are crucial for effectively training machine learning models without succumbing to overfitting. Additionally, data repositories foster collaboration and data sharing among researchers, accelerating scientific breakthroughs in understanding the intricate relationship between the gut microbiome and human health.

Achieving standardization in data collection, processing methods, and formats is paramount for advancing gut microbiome research (8). Modern data stewardship emphasizes the need for data to be Findable, Accessible, Interoperable, and Reusable (FAIR), a principle that is critical for the long-term value of microbiome repositories (20). The lack of standardized metadata and data structures remains a significant challenge, limiting the effective reuse of publicly available data for machine learning applications (21). Adopting FAIR principles not only facilitates data sharing and collaboration but also empowers new researchers to more readily enter the field. Furthermore, standardized methodologies ensure data consistency and comparability, allowing for more robust and reliable conclusions to be drawn from research studies.

Recent advancements highlight the critical synergy between metagenomics and culture-based research methods (9,10). While metagenomics offers a comprehensive overview of the microbial communities present in the gut, it may not always reveal the functional capabilities of individual species. On the other hand, culture-based methods enable the isolation and characterization of individual bacterial strains, providing deeper insights into their specific functions and interactions within the gut ecosystem. However, harnessing the full potential of this synergistic approach hinges on the development of robust and accessible database systems.

The expanding field of gut microbiome research generates a vast and complex data from both metagenomic and culture-based studies. This data encompasses diverse information, including microbial composition, functional profiles, strain-level characterization, and host health parameters. Effectively integrating and analyzing this diverse data is crucial for deciphering the intricate relationships between specific microbial communities and their impact on human health.

As research delves into the intricate complexities of the gut microbiome, the need for a scalable, accessible, and standardized data management infrastructure becomes increasingly crucial (10). Such infrastructure is necessary to effectively manage and analyze the complex data generated by diverse research methods. It needs to be scalable to accommodate the ever-growing volume of data, accessible to researchers across the globe, and standardized to ensure data consistency and facilitate seamless integration from diverse research efforts. By providing a central platform for storing, organizing, and sharing data generated from both metagenomics and culture-based studies, these systems enable researchers to:

- Correlate metagenomic data with functional information obtained from cultured isolates, leading to a more comprehensive understanding of the gut microbiome's functional potential.
- Identify novel bacterial species previously uncharacterized by metagenomic approaches, potentially unlocking new avenues for understanding the gut microbiome's role in health and disease.
- Facilitate collaborative research efforts by providing a shared resource for researchers across diverse disciplines, fostering innovation and accelerating scientific progress.

In summary, establishing an expansive, accessible data repository coupled with standardizing research methodologies, is a fundamental driver of scientific advancement. This approach holds the promise of surmounting the challenges posed by small sample sizes, fostering collaborations, and accelerating the pace of discovery in unraveling the intricate dynamics of the gut microbiome.

2. Microbiome Data Sharing and Database Structures

While advancements in sequencing technologies have revolutionized the study of the human microbiome, effectively sharing and utilizing this vast amount of data remains a significant challenge. Researchers currently rely on several existing avenues for data sharing, each with its own advantages and limitations:

- Public Databases: Platforms such as NCBI's SRA and ENA offer free data deposition but can be challenging to navigate and may not effectively accommodate all types of microbiome data (11,12). Additionally, these repositories primarily focus on raw sequencing data in the FASTQ format (13), which requires further processing before analysis.
- Specialized Databases: Entities such as the Human Microbiome Project (HMP) and the American Gut Project databases cater specifically to microbiome research and offer processed data in formats compatible with widely used analysis tools. However, these repositories might not encompass the breadth and diversity of data required for comprehensive research endeavors (1,14).
- Cloud-Based Platforms: Services like QIAGEN's CLC Genomics Cloud and BaseSpace Cloud offer convenience but might pose affordability issues for some researchers (15). However, associated costs might pose a barrier for some researchers, particularly those from low- and middle-income countries.

However, these platforms present several limitations that create challenges for researchers:

- Lack of Standardization: The absence of uniformity in data collection, processing, and storage methods across different platforms impedes effective comparison and integration of data from various studies (16). This is particularly problematic for researchers attempting to conduct meta-analyses or identify larger trends across diverse datasets.
- Data Silos: Fragmented data storage across numerous platforms creates data silos, where information remains inaccessible to researchers outside specific institutions or research groups. This hinders collaboration and slow down overall research progress (7).
- Complex Interfaces: User interfaces of some public databases can be challenging to navigate, especially for researchers with limited computational expertise. This could create a barrier to accessing and utilizing valuable data, hindering research efforts (16).
- Cost Barriers: The subscription fees associated with certain cloud-based platforms can restrict access for researchers from institutions with limited funding or from low- and middle-income settings. This creates an uneven playing field and limits global participation in microbiome research (16).

Challenges Faced by Researchers:

These limitations translate into concrete challenges for researchers. Disparate data storage makes it difficult to conduct comparative studies, while complex interfaces can significantly hamper the efficient utilization of available data. As a result, researchers often resort to scraping data from multiple sources, a process plagued by rate limits, CAPTCHA restrictions, and ethical concerns regarding data ownership and reuse. This hinders legitimate access to data and ultimately impedes research progress. In response to these challenges, the research community increasingly recognizes the need for a centralized, standardized, and user-friendly repository for microbiome data. Such a repository would streamline data sharing, enhance collaboration, and accelerate the pace of microbiome research by:

- Facilitating standardized data deposition and storage: Implementing consistent data formats and metadata standards would enable seamless data exchange and integration across diverse studies.
- Enhancing data accessibility and discoverability: A user-friendly interface with robust search and filtering options would empower researchers to easily locate and access relevant datasets, regardless of their technical expertise.

- Promoting data sharing and collaboration: A central repository would encourage researchers to share their data openly, fostering collaboration and accelerating scientific discovery.
- Reducing cost barriers: By eliminating the need for subscriptions to multiple platforms, a centralized repository could potentially provide open access to data, democratizing participation in microbiome research.

The development of such a repository holds immense importance to unlock the full potential of microbiome research, paving the way for groundbreaking discoveries in our understanding of human health and disease.

Call for a Centralized and User-Friendly Repository:

A centralized, standardized, and user-friendly repository for microbiome data is increasingly necessary. Such a repository would streamline data sharing, enhance collaboration, and drive the acceleration of microbiome research by mitigating the challenges encountered with current options.

2.1 Data Structures of Microbiome Databases

Understanding the intricacies of the microbiome necessitates the utilization of specialized databases housing copious amounts of sequencing data gathered from diverse studies. These databases structure and organize data in distinct formats, each catering to specific needs and functionalities.

National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA)

The NCBI SRA (11) primarily stores raw sequencing data, including microbiome data, in the widely used FASTQ format (13). Each FASTQ record encapsulates crucial information: a unique read ID, the DNA sequence itself, a quality score reflecting the read's quality, and metadata encompassing sample source, sequencing platform, and data processing steps (11). The metadata adheres to the Sequence Read Archive metadata format (SRA-STD), providing a standardized framework for describing sequencing data (11).

European Nucleotide Archive (ENA)

Similar to NCBI SRA, ENA (12) houses raw sequencing data, including microbiome data, in the FASTQ format (13). Its structure mirrors SRA's, utilizing the ENA-STD metadata format, extending beyond sequencing metadata to include data management and access-related information (12).

Human Microbiome Project (HMP) Database

Distinct from repositories storing raw data, the HMP database (1) hosts processed microbiome data. This encompasses OTU (Operational Taxonomic Unit) tables—a tally of distinct microbial species groups—and metagenomic profiles detailing gene abundance and functional pathways within a microbiome sample (17,18).

These formats align with widely used microbiome analysis tools and are complemented by comprehensive metadata encompassing sample demographics, clinical information, and data processing specifics (1).

American Gut Project Database

The American Gut Project database (14) distinguishes itself by storing processed microbiome data in the QIIME format (19), specifically designed for microbiome analysis. Alongside standardized metadata for sample demographics and collection dates, it incorporates dietary information, catering to unique research needs (14).

Comparative Overview

Each database employs the FASTQ format for raw sequence data and adopts standardized metadata practices. However, the divergence arises in their additional features and data processing capabilities. NCBI SRA and ENA emphasize standardized sequencing metadata, while the HMP database provides formats compatible with various

analysis tools and comprehensive supplementary metadata. The American Gut Project prioritizes dietary information alongside demographic details.

The following table summarizes the key differences in the data structures of the four microbiome databases:

Table 1. Comparison of selected databases.

Database	Data Format	Metadata Format	Additional Features
NCBI SRA	FASTQ	SRA-STD	Standardized metadata for sequencing data
ENA	FASTQ	ENA-STD	Standardized metadata for sequencing data and data management
HMP	OTU tables, metagenomic profiles	Standardized formats compatible with microbiome analysis tools	Additional metadata for sample demographics, clinical information, and data processing
American Gut Project	QIIME format	Standardized format for microbiome data analysis	Standardized metadata for sample demographics, collection date, and dietary information

2.1.1 Relevance in Microbiome Research: Interoperability and its Challenges

While the use of standardized formats like FASTQ and metadata across various databases facilitates a degree of interoperability, enabling data exchange and basic comparisons, achieving truly comprehensive cross-database analyses remains challenging. Several factors contribute to this complexity:

- Variations in Data Processing: Despite the standardization of formats, individual researchers and studies might employ diverse data processing pipelines involving unique filtering, normalization, and quality control steps. These variations can introduce subtle biases and inconsistencies that can complicate the meaningful comparison of data from different sources, potentially leading to misleading conclusions.
- Specialized Tools and Workflows: Different databases often cater to specific analysis pipelines and workflows, often relying on specialized software or analytical tools. This can create a barrier for researchers unfamiliar with these tools, hindering their ability to effectively analyze and compare data across diverse databases.
- Incompleteness of Metadata: While standardized metadata formats exist, the specific information captured, and its level of detail can vary across studies. This can make it challenging to fully understand the context and potential confounding factors associated with data from different sources, further limiting the reliability of cross-database comparisons.

These challenges highlight the crucial role that database structures play in shaping the landscape of microbiome research. Standardized formats and comprehensive metadata are essential for facilitating data accessibility and enabling basic comparisons. However, achieving truly meaningful integration and analysis across diverse studies necessitates further advancements in:

- Standardization of Data Processing Pipelines: Establishing consistent guidelines and best practices for data processing steps would minimize the introduction of biases and inconsistencies, allowing for more reliable comparisons across diverse datasets.
- Development of Interoperable Analysis Tools: Universal analysis tools compatible with data from various databases would empower researchers to seamlessly analyze and compare data regardless of its original platform or processing pipeline. This would encourage a more holistic approach to microbiome research, leveraging the collective wealth of data across diverse studies.
- Enhanced and Standardized Metadata Capture: Expanding the scope and detail of standardized metadata would provide researchers with a more comprehensive understanding of the context and potential confounding factors associated with each dataset, fostering more robust and reliable cross-study comparisons.

By addressing these challenges and promoting advancements in data standardization, interoperability, and analysis tools, we can unlock the full potential of cross-database comparisons in microbiome research. This will ultimately lead to a deeper understanding of the complex interactions within the human microbiome and its multifaceted impact on human health and disease.

2.1.2 Implications for Advancements: Striving for Standardization and Integration

The inherent dissimilarities in data structures across microbiome databases pose significant challenges for effective data sharing, analysis, and cross-study comparisons. These limitations hinder progress by compartmentalizing data, which prevents researchers from identifying larger trends or patterns that might only emerge from a more comprehensive perspective. Furthermore, the lack of standardization makes it difficult to conduct robust meta-analyses—a powerful research approach that integrates data from multiple studies to draw more definitive and generalizable conclusions about the microbiome's role in various contexts. Ultimately, these barriers to collaboration and knowledge sharing slow the overall pace of scientific discovery in the field of microbiome research.

To overcome these challenges and unlock the full potential of this research, it is imperative to strive for consistent standardized formats, inclusive metadata, and integrated analysis tools. This can be achieved by promoting the wider adoption of existing data standards, such as MIxS (Minimum Information for any (x) Sequence), which would facilitate consistent data collection and reporting, thereby promoting interoperability and simplifying data integration. Concurrently, developing user-friendly analysis tools capable of working seamlessly with data from diverse databases would empower all researchers, regardless of their technical expertise, to effectively analyze and compare data across platforms. Finally, establishing centralized data repositories with integrated analysis tools would foster open data sharing and collaboration, enabling the research community to leverage the collective power of diverse datasets. By pursuing these strategies, we can create a more integrated ecosystem that leads to improved data accessibility and enhanced collaboration, ultimately enabling us to unravel the intricate relationship between the human microbiome and human health with greater clarity and precision.

3. GMrepo as a Potential Solution

Existing limitations in microbiome data sharing have impeded collaborative research and hindered insights into the human gut microbiome (7,16). To address these challenges, GMrepo was developed—a curated database and platform tailored for storing, analyzing, and sharing human gut microbiome data. It stands out by providing a comprehensive, user-friendly, and accessible platform that surmounts the limitations of current options (7,16).

3.1 GMrepo's Unique Attributes: Overcoming Existing Challenges

GMrepo's strength lies in its adherence to established data standards, ensuring data consistency and interoperability. It simplifies data comparison and integration across diverse microbiome datasets, fostering meaningful conclusions (16). Moreover, GMrepo offers a user-friendly interface, democratizing data management and analysis for researchers of varying expertise levels (7,16).

Addressing Existing Limitations:

- Standardization: GMrepo adheres to established Genome Standards Consortium (GSC) standards for metagenomic data, ensuring data consistency and interoperability (16). This allows researchers to combine and compare data from diverse sources seamlessly, leading to more robust and reliable conclusions.
- Data Silos: By centralizing human gut metagenomes, GMrepo breaks down data silos, previously hindering accessibility and collaboration. This streamlined approach facilitates easy data access and sharing, fostering a more collaborative research environment (7).
- User-Friendly Interface: Recognizing the diverse technical skills of researchers, GMrepo prioritizes user-friendliness. Its intuitive interface empowers researchers of all experience levels to effectively manage and analyze data, democratizing participation in microbiome research (7,16).
- Cost Barrier: As an open-source and freely accessible platform, GMrepo eliminates cost barriers that might hinder access to valuable microbiome data, particularly for researchers from resource-limited settings (7,16).

Beyond mitigating existing limitations, GMrepo introduces further advantages:

- Comprehensive Metadata: GMrepo goes beyond storing data; it also furnishes exhaustive metadata for each dataset. This includes details like sample collection methods, sequencing techniques, and data processing steps. This rich metadata allows researchers to interpret data more effectively by providing crucial contextual information (16).
- Integrated Analysis Tools: Recognizing the value of integrated analysis, GMrepo offers a suite of built-in data analysis tools. These tools enable researchers to visualize data, identify differential abundance patterns, and conduct correlation analyses directly within the platform. This streamlines the research workflow and facilitates in-depth exploration (16).
- Collaboration Tools: GMrepo fosters research collaboration by providing dedicated features for data and analysis sharing. This allows researchers to collaborate seamlessly, share findings, and expedite research endeavors (7,16).

Despite its strengths, GMrepo could enhance its functionality by addressing certain areas:

- Batch Download: Implementing a feature for batch downloading of raw run data would aid researchers requiring deeper analyses.
- Expanded Audience: Broadening its scope to accommodate data analysts alongside healthcare professionals would enhance versatility.
- API Enhancement: Increasing API request limits would bolster capabilities for large-scale analyses.
- Refined Data Retrieval: Implementing batch download options based on specific parameters would expedite data retrieval for extensive projects.
- Longitudinal Tracking: Incorporating mechanisms to track changes in microbiome abundances over time would add value, albeit challenging due to typical data collection focusing on single-time-point samples.

3.2 GMRepo's Data Structure

GMrepo's data structure underpins its ability to efficiently manage, store, and analyze human gut microbiome data. Its adherence to established data standards ensures interoperability, enabling seamless data exchange between GMrepo and other sources (16).

Key Components:

- Raw Data Storage: GMrepo utilizes the FASTQ format to store raw microbiome sequencing data, accommodating data generated by diverse sequencing technologies (13). Each FASTQ record encapsulates crucial information, including DNA sequence reads, quality scores, and detailed metadata about the sample, sequencing specifics, and data processing steps (11).
- Processed Data Management: Beyond raw data, GMrepo also houses processed microbiome data in formats like OTU tables and metagenomic profiles (17,18). OTU tables provide taxonomic summaries, indicating the relative abundance of different microbial species groups within a sample. Metagenomic profiles, on the other hand, offer functional insights by detailing the genes and pathways present in the microbiome (17,18).

GMRepo's standardized formats streamline data exchange, ensuring compatibility with diverse analysis tools and platforms. Comprehensive metadata accompanying each dataset enriches researchers' understanding by providing contextual information about samples and processing steps (11).

Notable Features:

- Adherence to Data Standards: Follows Genome Standards Consortium (GSC) metagenomic data standards, ensuring data consistency and interoperability (16).
- Modular Organization: Divides its data structure into modules for raw and processed data, facilitating efficient management (16).

- Standardized Formats: Utilizes standardized formats for both raw and processed data, easing integration with external tools (16).
- Comprehensive Metadata: Accompanies datasets with comprehensive metadata, offering contextual insights (11).

In essence, GMRepo's data structure is meticulously designed to support efficient storage, management, and analysis of human gut microbiome data, ultimately facilitating data sharing, collaboration, and research advancements in microbiome science.

3.3 Conclusion

GMrepo emerges as an invaluable resource for microbiome researchers, offering a comprehensive and accessible platform for human gut microbiome data. With further development addressing identified limitations, it stands to evolve into a more potent tool, substantially contributing to our comprehension of the human gut microbiome's impact on health.

4. Architectural Blueprint for a Resilient Microbiome Data Platform

To address the challenges of data scalability, interoperability, and the need for advanced, privacy-preserving analytics, we propose a conceptual blueprint for a resilient microbiome data platform. The architecture is designed to be modular, scalable, and user-centric, providing a robust foundation for both current and future research needs. It moves beyond a simple data repository to create an integrated ecosystem for analysis and collaboration.

4.1 Conceptual Architecture

The proposed platform is built on a multi-layered, microservices-based architecture (Figure 1). This design separates concerns, enhances security, and allows for independent scaling of components to meet demand. The key layers are the Data Layer, the Processing & Analysis Layer, and the Application & Access Layer.

The AI / ML Service is designed to move beyond basic analytics and incorporate cutting-edge computational methods. The service will provide infrastructure to support advanced deep learning models - such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) which are increasingly used to identify complex patterns and hidden relationships within high-dimensional, sparse microbiome data (22). Furthermore, this service will facilitate the integration of multi-omics data (e.g., metagenomics, metaproteomics, metabolomics) to enable a more holistic understanding of microbial community function and to accelerate the discovery of robust clinical biomarkers (23).

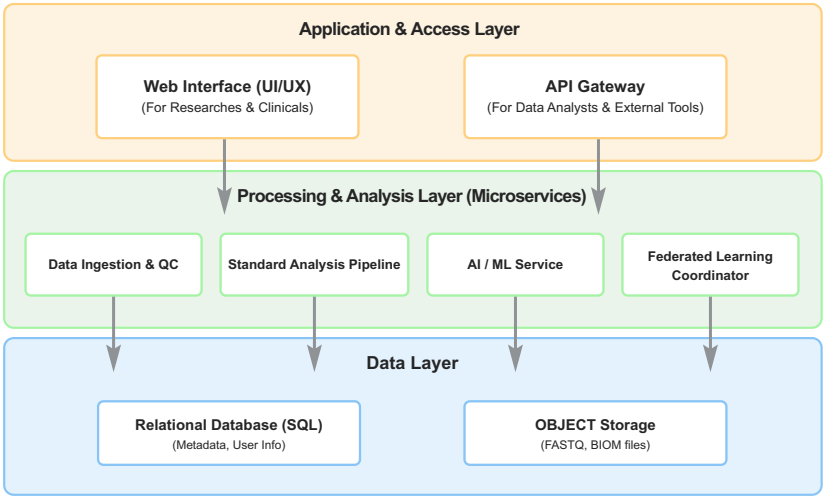

Figure 1. Conceptual architecture of the proposed platform.

4.2 Detailed Component Descriptions

Each component within the architectural layers is designed with specific roles and responsibilities to provide a clear overview of the platform's functionality.

In the Data Layer, the separation of storage is critical for efficiency and cost-effectiveness. The Relational Database (SQL) is optimized for structured, queryable data. It would house all sample metadata (e.g., subject_id, time_point, clinical variables), user account information, access control lists, and logs of analysis jobs. This allows for rapid querying, such as "find all samples from female subjects aged 30-40 with IBD." In contrast, the Object Storage (e.g., Amazon S3, Google Cloud Storage) is designed for large, unstructured binary files. It would store the raw FASTQ sequence files and large processed data files like BIOM tables and metagenomic assemblies. This two-pronged approach ensures that the database remains nimble for metadata queries while leveraging cost-effective, scalable storage for bulk data.

In the Processing & Analysis Layer, each microservice operates independently:

- The Data Ingestion & QC service acts as the gatekeeper. Upon data upload, it would automate a multi-step workflow: first, validating file integrity and format (e.g., confirming valid FASTQ syntax); second, running standardized quality control tools like FastQC to generate reports on sequence quality; and third, performing pre-processing steps such as adapter trimming (e.g., with Trimmomatic) and quality filtering to remove low-quality reads before the data is passed to the main storage and analysis pipelines.
- The Standard Analysis Pipeline provides baseline, reproducible analysis for all ingested data. This service would be configurable but could, for example, execute a QIIME 2 or Mothur workflow for 16S rRNA data to perform OTU clustering or generate ASVs (Amplicon Sequence Variants). For shotgun metagenomics data, it could run taxonomic classification (e.g., with Kraken 2) and functional profiling (e.g., with HUMAnN 3). The results would be stored and linked to the parent samples.
- The AI / ML Service, as previously noted, provides advanced analytics. Beyond supporting deep learning models, it would also offer tools specifically designed to handle the compositional nature of microbiome data (e.g., using Aitchison distance and centered log-ratio transformations) and to perform feature selection on high-dimensional data to identify the most salient microbial features for predictive modeling.
- The Federated Learning Coordinator is the orchestration engine for privacy-preserving analyses. It maintains a registry of participating institutions (nodes), manages the distribution of global models, and securely aggregates the returned model updates using techniques like Secure Aggregation to prevent inference on any single institution's contribution.

4.3 Technical Specifications for Scalability and Resilience

To ensure scalability and resilience, the proposed architecture is designed as a cloud-native platform utilizing a microservices framework. Each component (e.g., Data Ingestion, Analysis Engine) operates as an independent, containerized service. This allows for horizontal scaling using container orchestration technologies like Kubernetes, where resources for a specific service can be increased based on demand. This approach, widely adopted in modern data-intensive applications, allows the platform to handle growing data volumes and user loads efficiently while avoiding single points of failure, thereby ensuring high availability.

4.4 Privacy-Preserving Federated Analysis

A key innovation of this platform is its native support for federated learning (FL). This approach has been identified as a critical methodology for addressing the systemic and privacy challenges inherent in healthcare data (24). FL allows researchers to build robust machine learning models by training them across multiple decentralized datasets (e.g., at different hospitals) without centralizing sensitive patient data. The feasibility of such privacy-preserving federated learning (PPFL) has been successfully demonstrated in real-world applications using FAIR health data, proving that it is possible to achieve high predictive accuracy while maintaining data privacy and institutional control (25).

```
# Pseudo-code for a Federated Analysis Request
FUNCTION startFederatedAnalysis(analysisType, queryParameters):
  # 1. Researcher initiates analysis via the Web Interface.
  #    The request is sent to the Federated Learning Coordinator.
  COORDINATOR sends analysisModel and parameters to participating Institutions.

  # 2. Each institution's local system runs the model.
  #    The model is trained ONLY on the local, private data.
  FOR EACH institution IN participatingInstitutions:
    localModelUpdate = institution.runLocalTraining(analysisModel, localData)
    # 3. ONLY the anonymized model updates (e.g., weights) are returned.
    #    Raw patient data never leaves the institution.
    SEND localModelUpdate to COORDINATOR

  # 4. Coordinator aggregates the updates to build a global model.
  globalModel = COORDINATOR.aggregateUpdates(all_local_updates)

  # 5. Final, aggregated result is returned to the researcher.
  RETURN globalModel to Researcher
END FUNCTION
```

Figure 2. Pseudo-code for a federated analysis request.

5. Practical Applications and Sustainability

While the proposed platform is a conceptual framework, it is designed to address tangible, real-world research challenges. This section outlines a hypothetical use-case to illustrate its practical utility, details the methodology for longitudinal data management, and discusses the long-term vision for sustainability.

5.1 Hypothetical Use-Case Scenario: A Longitudinal Probiotic Study in IBD

Consider a multi-center clinical trial investigating the effect of a novel probiotic supplement on the gut microbiome of patients with Inflammatory Bowel Disease (IBD). Researchers aim to track microbial shifts over a six-month period and correlate them with clinical symptom scores.

- Data Ingestion and Standardization: Researchers from participating hospitals upload their raw sequencing data (FASTQ files) and associated clinical metadata (e.g., patient ID, diagnosis, symptom scores, time point) via the platform's secure web interface. The Data Ingestion & QC service automatically validates the data formats and runs quality control pipelines, ensuring all data entering the system is standardized.
- Longitudinal Analysis: Using the integrated analysis tools, a researcher initiates a time-series analysis. The platform leverages the stored time-point metadata to plot changes in alpha-diversity, beta-diversity, and the relative abundance of specific taxa (e.g., Faecalibacterium prausnitzii) for each patient over the six-month trial. The system can correlate these microbial shifts with the provided symptom scores, identifying potential time-dependent relationships.
- Federated Machine Learning for Biomarker Discovery: The research consortium wants to build a predictive model to identify baseline microbial signatures that predict a positive response to the probiotic. To increase statistical power, they decide to train their model against data from another, independent IBD cohort hosted at a non-partner institution. Using the Federated Learning Coordinator, they initiate a request. The analysis model is sent to the other institution, trained locally on their private data, and only the anonymous model parameters are returned. This allows the researchers to build a more robust and generalizable predictive model without ever compromising the privacy of the external cohort's data.

5.2 Hypothetical Use-Case Scenario 2: Public Health Surveillance of Antimicrobial Resistance

A national public health agency wants to monitor the prevalence of antimicrobial resistance (AMR) genes in gut microbiomes across the country.

- Continuous Data Aggregation: Hospitals and diagnostic labs across the country continuously upload shotgun metagenomic data from stool samples to the platform as part of routine diagnostics. The platform's ingestion service processes the data and runs a standardized AMR gene detection pipeline (e.g., using the CARD database).
- Geospatial and Temporal Analysis: Using the platform's interface, public health officials can create real-time dashboards that visualize the geospatial distribution of specific AMR genes (e.g., mcr-1, blaKPC). They can track the emergence of new resistance genes over time and set up alerts for when the prevalence of a high-risk gene exceeds a certain threshold in a specific region.
- Secure Cross-Jurisdictional Comparison: The agency wants to compare its AMR trends with a neighboring country's without sharing sensitive patient data. Through a federated query, they can ask: "What is the overall prevalence of carbapenemase-producing organisms in our respective populations?" Each jurisdiction runs the query on its own private data, and only the secure, aggregated result (the final prevalence number) is shared, enabling international collaboration while respecting data sovereignty.

5.3 Methodology for Longitudinal Data Management

The platform's ability to handle time-based tracking is a core architectural feature, not an afterthought. It is implemented through the database schema and analytical modules.

- Database Schema: The relational database includes specific fields within the metadata tables to support longitudinal studies. Every sample is linked to a unique subject_id and must have a time_point attribute (e.g., Day 0, Day 30, Day 90) and a collection_date. This structure ensures that all samples from a single individual can be easily grouped and ordered chronologically.
- Integrated Analysis Tools: The platform's built-in analysis tools are designed to recognize and utilize this temporal metadata. Time-series visualization tools will automatically plot data along the time axis, and statistical models like linear mixed-effects models will be pre-configured to handle repeated measures data, allowing researchers to properly account for inter-individual variability over time.

5.4 Sustainability and Long-Term Vision

The development and maintenance of such a platform represent a significant undertaking. A sustainable model is crucial for its long-term success and upkeep.

- Funding Model: A hybrid funding model is envisioned. Initial development could be supported by national and international research grants focused on scientific infrastructure. Long-term operational costs could be covered by a tiered institutional or enterprise subscription model, providing premium features (e.g., dedicated computational resources, advanced support) to larger organizations, while ensuring free access to core data exploration and analysis tools for academic researchers, thereby promoting equity.
- Governance and Upkeep: A scientific advisory board composed of experts in microbiology, bioinformatics, and data science would govern the platform to guide its development and ensure it continues to meet the evolving needs of the research community. Regular maintenance and updates would be managed by a dedicated technical team, with a focus on incorporating new analytical methods and maintaining security protocols.

6. Discussion: Catalyzing Innovation with a Benchmark Microbiome Dataset

The rapid advancement of fields like machine vision was significantly accelerated by the creation of benchmark datasets, with the MNIST database of handwritten digits serving as a prime example. While the image data itself was simple, its widespread availability in a standardized format with clear labeling created a common ground

for the global research community. This lowered the barrier to entry for computational scientists and, crucially, enabled the direct, objective comparison of novel algorithms, which in turn spurred immense innovation.

A similar paradigm-shifting opportunity exists within microbiome research. The creation of a benchmark "Microbiome Reference Dataset," built upon FAIR principles, could similarly democratize research and foster a new wave of computational method development. It is, however, crucial to acknowledge the fundamental differences in complexity and structure; microbiome data is inherently more sparse, compositionally complex, and deeply dependent on extensive, standardized metadata for proper interpretation. Therefore, a direct analogy regarding data structure is not the intention. Instead, the critical lesson from MNIST is its powerful catalytic effect on an entire research ecosystem.

The primary obstacle to creating such a resource is not a lack of data, but the persistence of the exact challenges our proposed platform is designed to solve: data fragmentation, non-standardized formats, and siloed access. The architectural blueprint detailed in this paper provides the necessary foundation to build, curate, and disseminate such a high-value dataset. The platform's key features are direct enablers of this vision. Standardized Data Ingestion and Processing ensures that all data within the benchmark set is consistent and comparable. Rich Metadata Association adheres to community standards to provide the deep contextual information essential for microbiome analysis. Finally, API-driven Access provides the simple, programmatic access that made MNIST a staple for computational researchers.

A publicly available, well-documented benchmark dataset enabled by this platform would have profound implications for the field. It could be used to validate new biomarkers, develop robust disease classification models using advanced AI, and create a standardized testbed for novel analytical tools. In conclusion, while not a direct parallel in data complexity, the MNIST dataset serves as a powerful testament to how a community-wide, accessible resource can fuel progress. By providing the infrastructure to create a similar resource for the microbiome community, our proposed platform aims to catalyze the next generation of discovery in understanding the microbiome's role in human health.

7. Conclusion

In the evolving landscape of microbiome science, the capacity to effectively manage, integrate, and analyze vast datasets is paramount to translating research into clinical impact. This paper addressed the persistent challenges that have led to a "data-rich, information-poor" paradox in the field—namely, data fragmentation across disparate silos, a lack of methodological standardization, and barriers to privacy-preserving analysis. In response, we have proposed a comprehensive architectural blueprint for a next-generation microbiome data platform. Our model transcends the role of a simple repository by creating an integrated ecosystem that supports sophisticated, longitudinal analysis while upholding the highest standards of data stewardship through FAIR principles.

The core of our proposal is a flexible, multi-layered architecture that ensures scalability and facilitates the integration of advanced computational tools. By incorporating native support for federated learning (FL) and AI-driven analytics, the platform is designed to empower researchers to ask more complex questions of their data. This approach not only enables large-scale, cross-institutional studies that were previously infeasible due to privacy constraints but also helps bridge the gap between bioinformatics specialists and clinical researchers by providing user-friendly access to powerful analytical pipelines. The potential impact is the acceleration of discovery, leading to more robust biomarkers and a deeper, more mechanistic understanding of microbial dynamics in health and disease.

The conceptual framework presented here provides a clear roadmap for future development. The critical next phase involves the creation of a functional prototype, beginning with the Data Ingestion & QC and Longitudinal Analysis modules. Validating the proposed design will require not only technical performance benchmarking but also user-centric evaluation through pilot programs with academic and clinical partners. These collaborations will be crucial for refining the platform's features and ensuring they meet real-world needs. In parallel, a key long-term goal is the cultivation of a sustainable community, supported by a hybrid funding model and transparent governance. This includes establishing clear ethical guidelines for data use within the federated network and actively engaging with international standards bodies like the Global Alliance for Genomics and Health (GA4GH) to ensure continued interoperability.

By building this foundation, we can move closer to an era where the full potential of microbiome data is unlocked. An infrastructure of this nature is a critical enabler for the future of medicine, paving the way for highly personalized therapeutic interventions, proactive public health strategies, and ultimately, transformative improvements in human health worldwide.

8. Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval and consent to participate

Not applicable. This study is a conceptual proposition and did not involve human participants, human data, human tissue, or animals.

Consent for publication

Not applicable.

Data availability

All necessary data supporting the findings of this study are available within the manuscript.

Materials availability

All necessary materials are described within the manuscript.

Code availability

Not applicable. This study presents a conceptual framework; no code was generated or used.

Author contribution

T. Ráčil conceived the study concept, wrote the main manuscript text, and prepared the formatting. A. Štefek and P. František reviewed the manuscript and contributed to refining the conceptual framework. All authors have read and approved of the final manuscript.

9. References

1. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207-214. doi: 10.1038/nature11234.
2. Cryan JF, Dinan TG. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. Nat Rev Neurosci. 2012;13(10):701-712. doi: 10.1038/nrn3346.
3. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016;14(8):e1002533. doi: 10.1371/journal.pbio.1002533.
4. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project. Nature. 2007;449(7164):804-810. doi: 10.1038/nature06244.
5. Li W, Chen X, Xie L, et al. Bioelectrochemical systems for groundwater remediation: The development trend and research front revealed by bibliometric analysis. Water (Basel). 2019;11(8):1532. doi: 10.3390/w11081532.

6. James G, Witten D, Hastie T, et al. An introduction to statistical learning. New York: Springer; 2013. 426 p. doi: 10.1007/978-1-0716-1418-1
7. Gonzalez A, Knight R. Advancing analytical algorithms and pipelines for billions of microbial sequences. Nat Rev Microbiol. 2012;10(2):64-71. doi: 10.1016/j.copbio.2011.11.028
8. Bry F, Kröger P. A Computational Biology Database Digest: Data, Data Analysis, and Data Management. Distrib Parallel Databases. 2003;13(1):7–42. doi: 10.1023/A:1021540705916.
9. Ankrah NYD, Bernstein DB, Biggs M, et al. Enhancing Microbiome Research through Genome-Scale Metabolic Modeling. mSystems. 2021;6(6):e0059921. doi: 10.1128/mSystems.00599-21.
10. Hitch TCA, Afrizal A, Riedel T, et al. Recent advances in culture-based gut microbiome research. Int J Med Microbiol. 2021;311(3):151485. doi: 10.1016/j.ijmm.2021.151485.
11. NCBI Resource Coordinators. Sequence Read Archive (SRA) [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); [cited 2024 Aug 22]. Available from: https://www.ncbi.nlm.nih.gov/sra
12. EMBL-EBI. European Nucleotide Archive [Internet]. Hinxton (UK): European Molecular Biology Laboratory - European Bioinformatics Institute; [cited 2024 Aug 22]. Available from: https://www.ebi.ac.uk/ena
13. Cock PJA, Fields CJ, Goto N, et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767-1771. doi: 10.1093/nar/gkp1137.
14. McDonald D, Hyde E, Debelius JW, et al. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems. 2018;3(3):e00031-18. doi: 10.1128/msystems.00031-18.
15. QIAGEN. QIAGEN CLC genomics cloud engine [Internet]. Hilden (DE): QIAGEN; [cited 2024 Aug 22]. Available from: https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/enterprise-ngs-solutions/qiagen-clc-genomics-cloud-engine/
16. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174-180. doi: 10.1038/nature09944.
17. Yilmaz P, Parfrey LW, Yarza P, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 2014;42(Database issue):643-648. doi: 10.1093/nar/gkt1209.
18. Eren AM, Maignien L, Miller EL, et al. Minimum Entropy Decomposition: Unsupervised Clustering for Variable-Length Reads. Nat Methods. 2015 Jul;12(7):641-643. doi: 10.1038/nmeth.3368.
19. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335-336. doi: 10.1038/nmeth.f.303.
20. Dorst M, Zeevenhooven N, Wilding R, et al. FAIR compliant database development for human microbiome data samples. Front Cell Infect Microbiol. 2024;14:1384809. doi: 10.3389/fcimb.2024.1384809. PMID: 38774631; PMCID: PMC11106358.
21. Kumar B, Lorusso E, Fosso B, et al. A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. Front Microbiol. 202413;15:1343572. doi: 10.3389/fmicb.2024.1343572. PMID: 38419630; PMCID: PMC10900530.
22. Przymus P, Rykaczewski K, Martín-Segura A, et al. Deep learning in microbiome analysis: a comprehensive review of neural network models. Front Microbiol. 2025;15:1516667. doi: 10.3389/fmicb.2024.1516667. PMID: 39911715; PMCID: PMC11794229.
23. Dakal TC, Xu C, Kumar A. Advanced computational tools, artificial intelligence and machine-learning approaches in gut microbiota and biomarker identification. Front Med Technol. 2025;6:1434799. doi: 10.3389/fmedt.2024.1434799. PMID: 40303946; PMCID: PMC12037385.
24. Zhang F, Kreuter D, Chen Y, et al. Recent methodological advances in federated learning for healthcare. Patterns (N Y). 2024;5(6):101006. doi: 10.1016/j.patter.2024.101006. PMID: 39005485; PMCID: PMC11240178.
25. Sinaci AA, Gencturk M, Alvarez-Romero C, et al. Privacy-preserving federated machine learning on FAIR health data: A real-world application. Comput Struct Biotechnol J. 2024;24:136-145. doi: 10.1016/j.csbj.2024.02.014. PMID: 38434250; PMCID: PMC10904920.