International Journal of Data Science and Analytics (2026) 22:134
https://doi.org/10.1007/s41060-026-01107-8

REGULAR PAPER

Integrating microbiome data visualization into FAIRDatabase using edge functions

Roman van Eldijk[1] · **Shivam Kumar**[1] · **Vivek M Sheraton**[1]

Received: 29 January 2026 / Accepted: 23 March 2026

Abstract
Microbiome research continues to grow, so does the volume of data it produces. Yet privacy constraints on human-associated samples and the compositional nature of sequencing outputs make quick exploratory analysis difficult. This study extends the FAIRDatabase, an open-source, privacy-compliant infrastructure for microbiome data, with a visualization module designed to tackle both challenges. The module performs composition-aware beta diversity analysis using centered log-ratio transformation and the Aitchison distance metric. All computations are run within Supabase edge functions, which makes sure that sensitive data never leave the secure environment. To guide the design, requirements were derived from prior work and literature, covering compositional data analysis, beta diversity visualization, and principles for clear data interpretation. The resulting tool supports interactive heatmaps and principal coordinates analysis (PCoA) plots, with options for metadata-based coloring, variance explained labels, and color palettes chosen for accessibility and interpretability. In order to evaluate the module, ten participants have performed tasks on the module and filled in the system usability scale (SUS) questionnaire, which resulted in a mean SUS score of 86.8. Three of whom have been interviewed. They valued being able to quickly explore data without downloading files or facing contractual obstacles. Overall, this work shows that edge functions can support composition-aware microbiome analysis without compromising data security. It offers a starting point for building privacy-preserving visualization tools in research areas where data sensitivity is a significant concern.

Keywords Microbiome data · Compositional data analysis · Edge functions · Beta diversity · FAIR principles · Supabase

1 Introduction

The volume of microbiome sequencing data has increased rapidly in the last two decades [24]. "Microbiome" refers to the entire microbial ecosystem, namely the community of microorganisms (bacteria, fungi, and viruses) along with their "theater of activity", such as genes, metabolites and the surrounding environmental context [6, 7]. The field of microbiome research made great improvements in the early 2000s, and advances in high-throughput sequencing have enabled the large-scale study of microbial communities [26]. However, in the last decade the number of published datasets has increased so much that the research field has difficulty handling the amount of data coming in [24]. Inconsistent data formatting and insufficient metadata make it difficult to compare and reproduce results between studies [23]. This limited approach led to reproducibility and standardization issues [45].

In addition to the issues of standardization, human-associated microbiome data introduces a more fundamental challenge: privacy. Microbiome profiles are highly individual-specific and individuals could be uniquely identified in a population, highlighting the strict privacy constraints when the data concern human subjects [17]. Cho [12] showed that an individual can be revealed based on the human microbiome sample, violating GDPR compliance. These constraints significantly limit open sharing, which directly conflicts with key FAIR principles (Findability, Accessibility, Interoperability, and Reusability) that attempt to overcome data management issues through standardization and transparency [47]. As highlighted by Dorst et al. [14], the central

✉ Vivek M Sheraton
v.s.muniraj@uva.nl

Roman van Eldijk
roman.van.eldijk@student.uva.nl

Shivam Kumar
s.kumar@uva.nl

[1] Informatics Institute, Universiteit van Amsterdam, Amsterdam, The Netherlands

Published online: 13 April 2026

barrier to reusability is therefore not the volume of data but the inability to store, access and reuse human microbiome datasets without violating privacy laws.

To address these challenges, the FAIRDatabase was developed as an open-source infrastructure that applies the FAIR principles and integrates privacy-compliant strategies for human microbiome data. The FAIRDatabase provides standardized storage for data and suitable metadata and an interface that allows researchers to explore datasets without directly revealing sensitive raw information [14]. The FAIRDatabase deploys the FAIR principles by converting files to the right format, scanning the uploaded file type, raising an error if the file is not FAIR adherent, and by creating metadata about the dataset to make it interoperable [14].

Visualization is one of the first steps in exploring a new dataset [40], and tool-driven, discovery-driven research can lead to a deeper understanding of the data [38]. To enable visualization while maintaining privacy and security guaranties, edge functions play an important role. Edge functions process data close to the node rather than transmitting data to external servers, which results in a secure way of handling sensitive data and reduces latency [10]. This is essential for robust and reliable data handling in the healthcare sector [25]. Due to the strict privacy requirements of human microbiome data, edge functions are the necessary architectural choice to ensure data minimization.

This research addresses these gaps by extending the FAIRDatabase visualization module with composition-aware beta diversity analysis. This means that researchers exploring microbiome datasets may view incorrect visualizations, which can lead to misinterpretations of sample relationships. The edge function will be used to ensure reduced latency and to keep sensitive data in a secure environment. The visualizations will follow design principles that support better interpretability.

2 Methods

2.1 Compositional data analysis

2.1.1 The nature of compositional data

Compositional data consist of vectors of positive components that carry only relative information, typically constrained to sum to a constant such as 1 or 100% [36]. The sample space of compositional data is the simplex, defined as the set of positive vectors whose components sum to a constant [1]. This constant-sum constraint introduces a dependence between components: an increase in one component necessarily implies a decrease in at least one other, regardless of any true underlying relationship [36].

These properties distinguish compositional data from unconstrained multivariate data, which makes standard statistical techniques less suitable. Conventional methods assume that variables can vary freely across the real number line, but the variables are restricted to positive values within a bounded sum. Pearson first identified that analyzing raw compositional components with standard correlation measures produces spurious relationships [37]. Any meaningful analysis of compositional data must, therefore, consider their inherent relative nature and simplex geometry.

2.1.2 Microbiome data as compositional data

High-throughput sequencing is the primary method for generating microbiome data, since it is the optimal method for describing microbial composition and function [41]. In amplicon sequencing, a conserved marker gene, most commonly the 16 S rRNA gene for bacteria, is extracted from a sample and amplified using Polymerase Chain Reaction (PCR) [11]. The sequences are then clustered into Operational Taxonomic Units (OTUs) based on sequence similarity, typically at a 97% threshold, producing a count table that records how many reads were assigned to each OTU per sample [15]. However, this process produces inherently compositional observations. Sequencing instruments have a fixed capacity and are unable to capture the actual microbial load in the sample [18]. Consequently, sequencing data reflect only relative abundances; absolute counts of any taxon cannot be determined from data alone [33].

2.1.3 Log-ratio transformations

The solution to compositional constraints lies in log-ratio transformations, which capture relationships between components without being affected by the constant-sum property [1]. Because compositional data carry relative rather than absolute information, only the ratios between components are mathematically meaningful [20]. Taking the logarithm of these ratios converts the data from a multiplicative scale to an additive scale, enabling the application of standard statistical and visualization methods, with the centered log-ratio being the most essential [2, 20].

The centered log-ratio (CLR) transformation is defined as:

$$\mathrm{CLR}(x_i) = \ln\left(\frac{x_i}{g(x)}\right) \tag{1}$$

where x_i is the abundance of taxon i and $g(x)$ is the geometric mean of all components in the sample. The CLR transformation maps compositional data from the simplex to a dimensional subspace of Euclidean space, where distances and variances can be interpreted in the standard way [2].

A practical challenge arises from the frequency of zeros in microbiome data. Sequencing data are typically sparse, with up to 90% of entries being zero due to taxa that are absent or below detection limits [35]. Since logarithms are undefined for zero, various strategies have been developed to handle this issue. The most straightforward approach adds a small pseudo-count to all values before transformation [30]. Pseudocounts are often used in CLR and can vary from very low positive numbers, such as 10^{-8} to 0.01 [4, 21].

2.2 Beta diversity analysis

2.2.1 Measuring differences between communities

Beta diversity quantifies how microbial communities differ from one another across samples, sites, or experimental conditions [28], whereas alpha diversity describes the diversity within a single sample, beta diversity captures the diversity of species between samples to find clusters or similar samples [38]. Beta diversity analysis typically involves two steps: calculating pairwise distances between all samples and then visualizing the resulting distance matrix.

2.2.2 Distance metrics

The choice of distance metric determines what aspects of community difference are captured. Several metrics are commonly used in microbiome research, each with different properties and assumptions, such as Bray-Curtis, UniFrac, and Jenson-Shannon [38]. However, these distance metrics operate on relative abundance data and do not account for the compositional constraint. The Aitchison distance addresses the compositional constraints by calculating the Euclidean distance on the CLR-transformed data [2, 20], which first removes the compositional constraint through log-ratio transformation, then applies standard Euclidean distance calculation in the transformed space. For compositional microbiome data, the Aitchison distance provides a correct measure of dissimilarity that reflects genuine differences between samples.

2.2.3 Visualization methods for beta diversity

In order to visualize the beta diversity, we calculate the distance or dissimilarity matrix, which records distances between all samples, as we discussed in the previous section. The way this matrix is translated into visual representation determines what patterns researchers can detect and how easily the results can be interpreted [38].

Heatmaps provide a direct visualization of the distance matrix itself, capturing all data in a static manner and encoding (dis)similarity values as color intensity where both rows and columns represent samples, as shown in the paper by Lei et al. [27]. This approach preserves all relationships without dimensional reduction, allowing researchers to identify which specific sample pairs are most similar or dissimilar. Heatmaps work well for datasets with moderate numbers of samples but become difficult to read as sample counts increase, since the number of cells grows with sample size [38]. A solution to this is clustering the output into groups, as used in the paper of Lei et al. citeLei2017. Researchers frequently seek to uncover underlying patterns beyond mere comparisons of similar samples, with ordination-based methods being more popular in the literature [38].

Ordination methods reduce high-dimensional distance matrices to two or three dimensions for visualization, enabling researchers to identify clustering patterns and gradients that would be invisible in the raw matrix. Principal Coordinates Analysis (PCoA) has seen an upward trend in microbiome analyses and is a common ordination method in microbiome research [8]. PCoA takes a distance matrix as input and finds a representation that preserves the original distances as accurately as possible. The axes of a PCoA plot are positioned by the amount of variance they explain, with the first axis capturing the largest source of variation.

Non-metric Multidimensional Scaling (NMDS) offers an alternative to PCoA that prioritizes rank order preservation over exact distance reproduction [42]. NMDS iteratively adjusts point positions to minimize stress between the rank order of original distances and the rank order of distances in the reduced space. This makes NMDS more robust to nonlinear relationships but also means that axis scales are arbitrary and cannot be directly compared across plots [3].

The compositional biplot deserves special attention for microbiome data because it simultaneously displays both samples and taxa in the same plot [19]. After CLR transformation, standard PCA produces a biplot where sample scores and taxon loadings share a common coordinate system. This allows researchers to identify which taxa drive the separation between sample groups and to detect associations among taxa, making the biplot particularly valuable for hypothesis generation.

2.3 Existing visualization tools

To position the developed module within the landscape of existing microbiome visualization tools, a qualitative feature comparison was conducted. The comparison focuses on four criteria relevant to the research objectives: (1) support for composition-aware analysis, (2) privacy preservation through server-side computation without requiring data download, (3) accessibility for users without programming expertise, and (4) interactive visualization capabilities. QIIME 2 is a comprehensive platform that supports compositional analysis through the DEICODE plugin, which implements robust Aitchison PCA, but requires command-

line expertise and local data handling [16]. Phyloseq offers extensive visualization capabilities within R and CLR transformation is available through the companion microbiome package, but both require programming proficiency [31]. MicrobiomeAnalyst provides a user-friendly web interface with composition-aware options, yet processes data on external cloud servers, raising concerns for sensitive human microbiome data [13, 29]. Similarly, Calypso offers a web-based interface with various visualization types but also processes data on remote servers [48]. EzMAP provides an integrated pipeline with explicit support for beta diversity analysis, but requires data upload to external infrastructure [43]. The FAIRDatabase visualization module addresses the gap at the intersection of these criteria: it provides composition-aware beta diversity analysis through an accessible web interface while ensuring that sensitive data remain within the secure environment through edge function computation. A direct empirical comparison of latency and usability across platforms was not feasible within the scope of this study, as this would require equivalent evaluation protocols and identical datasets across all tools.

2.4 Edge functions

Processing sensitive microbiome data while maintaining privacy compliance requires cautious architectural decisions about where computation occurs. In traditional web applications, data is often transmitted to the client browser for processing, which poses risks when working with human microbiome data. Another approach is to perform computations on the server-side, transmitting only the results to the client. Supabase, the platform underlying FAIRDatabase, provides edge functions, which are serverless TypeScript functions that run on globally distributed infrastructure [44]. Edge computing traditionally refers to computational architectures that process data at or near the source rather than transmitting it to centralized servers [10]. Supabase Edge Functions operate on Deno Deploy infrastructure and should be serverless functions with direct database access, rather than processing data at the device or network edge. For the visualization module, edge functions offer several benefits. Edge functions are globally distributed, so the function runs on a regionally-distributed Edge Runtime node closest to the user for minimal latency [44]. Another benefit is that the calculation is performed within the function, and only the derived results are returned, which leads to data minimization[44]. Edge functions do also have limitations. They are designed for short-lived, stateless operations, and the documentation warns that long-running heavy computations should be moved to background workers [44]. However, for quick exploration tasks the computational load should be manageable.

2.5 Design science research

This research follows an adapted Design Science Research (DSR) approach [22]. While traditional DSR consists of six distinct activities [39], this study integrates six distinct activities: the introduction serves as problem identification, requirements elicitation defines solution objectives, implementation combines design/development with demonstration as the working prototype demonstrating that the module is indeed working, evaluation assesses the artifact with domain experts, and this study provides communication in the form of a discussion and conclusion.

2.5.1 Requirements elicitation

A focused literature search was conducted to identify papers directly informing composition-aware transformations, beta diversity visualizations and edge functionality. The goal was not to perform a systematic review, but to gather sufficient evidence to formulate a grounded set of requirements.

Sources were selected based on their relevance to the research question and scope, and requirements were extracted by identifying recommended practices. Deriving requirements from related work aligns with the Design Science Research framework, which holds that design artifacts should originate from an existing knowledge base of theories, prior research, and documented solutions to ensure the rigor of the research [22].

Requirements were then classified as functional or non-functional following Wiegers and Beatty [46]. A functional requirement is "a description of a behavior that a system will exhibit under specific conditions," while a non-functional requirement is "a description of a property or characteristic that a system must exhibit or a constraint that it must respect."

To structure implementation, requirements were prioritized using the MoSCoW method [32]. "Must have" requirements are essential for a functioning composition-aware beta diversity visualization, "should have" requirements significantly enhance usability or interpretation, "could have" requirements add value but are desirable but not critical, and "won't have" requirements are identified as out of scope for this project but are documented for future development. This prioritization ensures that development efforts focus on the most critical functionality within the research timeline.

2.5.2 System design and implementation

Following the collection and prioritization of requirements, the composition-aware beta diversity visualization module is developed as an extension to the existing FAIRDatabase infrastructure, keeping the original design choices in mind.

The module is designed to perform all computations on the data server-side via Supabase edge functions, ensuring that

sensitive microbiome data remain within the secure FAIRDatabase environment. The usage of edge functions divides the computation and storage within Supabase. PostgreSQL handles the storage and the edge function perform the calculations. The pipeline within the Supabase environment consists of three stages: Edge function requests data from the database using SQL queries, data gets transformed and calculations are performed on the data, and the produced results are sent to the frontend. The edge functions are implemented in TypeScript using the Supabase edge function framework. After implementation, the latency of the edge functions will be tested. The frontend of the visualization module is developed using JavaScript with Plotly.js.

The code for the visualization module can be found in the following GitHub repositories: https://github.com/romanvaneldijk/FAIRDatabase, https://github.com/SheratonMV/FAIRDatabase

2.5.3 Evaluation

The evaluation assesses whether the developed module enables researchers to complete beta diversity exploration tasks effectively, addressing the third sub-research question. Ten domain experts were recruited from microbiome research groups at the University of Amsterdam and affiliated institutions. Participants have experience with microbiome data research and have used analysis tools, but do not have prior experience with FAIRDatabase specifically. This selection ensures that feedback reflects the perspective of the intended users. All ten participants completed the System Usability Scale (SUS) questionnaire after using the module. Of these, three participants also completed task-based evaluation sessions with semi-structured interviews to gather in-depth qualitative feedback.

For the three interviewed participants, the evaluation proceeded as follows. First the participant is instructed on how the evaluation will take place and their consent for recording is asked. Then an introduction of the module will be given with the emphasis on the compositional nature of the data, beta diversity analysis and their specific visualizations, and the objective of this research. The participants will be asked general questions about their background and their affinity with microbiome data and they will be asked to complete a series of tasks within the visualization module. Tasks are designed to reflect realistic exploratory analysis scenarios and to probe both usability and interpretive accuracy. Example tasks include:

- How many datasets are available, and which dataset has the most samples?
- Explore the visualization through its functions. Which samples are the most similar?
- What transformation has been applied to the data? Where is this information shown?
- Change the visualization so that samples are colored by the 'Treatment' variable.

For each task, participants will be asked to verbalize their reasoning (think-aloud protocol).

After the completion of the tasks, the participant will be interviewed to gather information about their experience. The interview will be semi-structured with these main topics: perceived usefulness, usability and interaction, composition-aware features, suggestions and improvements.

Evaluation sessions were conducted under two different conditions due to practical constraints. Two sessions were conducted in person, while one session was conducted remotely via Zoom using the remote control feature, which allowed the participant to interact directly with the prototype interface. The remote session followed the same protocol as the in-person sessions to maintain consistency.

Additionally, two of the three participants were Dutch-speaking, and their evaluation sessions were conducted in Dutch to ensure participants could express their thoughts and feedback more naturally and precisely. Interview responses from these sessions were translated afterward to English for analysis. The third participant was evaluated in English.

Two types of data will be collected during the evaluation sessions. Quantitative data will be gathered through the System Usability Scale questionnaire, where participants rate ten statements on a five-point Likert scale ranging from strongly disagree to strongly agree. Qualitative data will be collected through the semi-structured interview responses addressing perceived usefulness, usability, composition-aware features, and suggestions for improvement. All sessions will be audio recorded to enable accurate transcription.

The SUS questionnaire responses will be scored using the standard calculation method, which yields a usability score between 0 and 100 for each participant [9]. To interpret the score, an adjective rating scale is used, ranging: "worst imaginable," "poor," "ok," "good," "excellent," and "best imaginable" [5]. Since the sample size is small, individual scores will be reported next to the mean, and results will be interpreted indicative rather than generalizable. Qualitative data interview transcripts will be analyzed using thematic analysis. Responses will be coded to identify recurring patterns, which will then be grouped into themes.

3 Results

3.1 Functional and non-functional requirements

In Table 1 the Functional requirements are displayed and in Table 2 the Non-functional requirements.

Table 1 Final set of functional requirements for the compositional microbiome visualization tool. Requirements were synthesized from previous work and related work analysis, then prioritized using the MoSCoW framework based on their relevance to compositional data analysis principles and user needs

REQ ID	Requirement	MoSCoW
F-001	The tool must provide the user with different options to handle zero values	Must have
F-004	The tool must be able to visualize the beta diversity metrics using ordination methods	Must have
F-005	The tool must be able to represent the distance matrices using heatmaps, with color intensity as dissimilarity score	Must have
F-010	The tool must allow users to select samples for ordination plots, with PCoA and NMDS being the most important plots	Must have
F-014	The tool must use color maps that vary in luminance when visualizing numerical values	Must have
F-015	The tool must improve accessibility by providing primarily color-blind-friendly color maps	Should have
F-016	The tool must include interactive PCA or PCoA plots that are colored by metadata variables or by groups	Should have
F-017	The tool must use interactive features like tooltips in heatmaps to provide more detailed information when data points are hovered over	Should have
F-019	The tool must be able to compare variables from the metadata to allow the user to explore correlations	Should have
F-101	A log-ratio transformation should be used to deal with the composite nature of the data, with centered log-ratio being the most important	Must have
F-103	The Aitchison distance metric, with a CLR transformation should be used for beta diversity analysis	Must have
F-106	Compositional Biplot is an important visualization method	Could have
F-107	Ordination plot axis should display the variance captured	Must have
F-108	Edge processing should be used to ensure the privacy of the used data	Must have

Table 2 Final set of non-functional requirements for the compositional microbiome visualization tool. Requirements were synthesized from previous work and related work analysis, then prioritized using the MoSCoW framework based on their relevance to compositional data analysis principles and user needs

REQ ID	Requirement	MoSCoW
NF-001	The tool should improve reproducibility by allowing users to access the complete code used	Could have
NF-002	The tool should promote understanding of the study by displaying metadata	Could have
NF-003	The tool should allow users to analyze multiple variables at the same time for data exploration	Could have
NF-004	The tool should create visualizations that maintain clarity, readability, and interpretability, even when handling large or complex datasets	Should have
NF-101	Use spatial position to visualize quantitative data	Must have
NF-102	Use spatial region and color hue for categorical data	Could have
NF-103	The visual encodings should display all the relevant information	Must have
NF-104	Coloring that suggests ordering of categorical data should be avoided	Should have

3.2 Data visualization module

The visualization module is implemented within FAIR-Database's existing infrastructure, which combines a Supabase-managed PostgreSQL database and a frontend built using HTML, CSS and JavaScript. The overall architecture follows a model in which all computation and processing is handled server-side through Supabase edge functions. User interactions in the browser trigger requests to these edge functions, which retrieve the necessary data, perform all transformations and calculations, and return only derived results for visualization. The latency results are shown in Fig. 1. Ten iterations were performed and for each iteration the processing time and network latency were measured.

The latency evaluation demonstrates that edge functions provide acceptable performance for exploratory analysis, with response times remaining below two seconds for datasets up to 80 samples and 10,000 OTUs. To contextual-

ize these results, a baseline comparison was conducted using scikit-bio (version 0.7.2) to perform the same CLR transformation, Aitchison distance calculation, and PCoA analysis locally on identical synthetic datasets, bypassing database access and network overhead. The baseline computation times ranged from 0.3 ms to 14 ms depending on dataset dimensions, compared to 67 ms to 1,596 ms for the edge functions. This overhead, averaging approximately 100–200 times slower than local computation, reflects the combined cost of network latency, database retrieval, and secure environment execution. This overhead represents the necessary trade-off for maintaining data privacy: while local computation would be faster, it would require transferring sensitive microbiome data outside the secure environment. Practical resource limits exist due to Supabase edge function constraints, which include a maximum memory of 256 MB and a CPU time limit of 2 s per request. In testing, a dataset with 90 samples and 10,000 OTUs exceeded these limits and failed to process, indicating that the current implementation is best suited for small to moderately sized datasets typical of exploratory analysis.

When a user requests a visualization, the workflow proceeds as follows. First, the user selects a dataset and its parameters in the web interface. These options are sent from the frontend to the appropriate Supabase edge function. The edge function then queries the PostgreSQL database for the requested abundance data and performs all required processing steps, including pseudocount addition, centered log-ratio (CLR) transformation, distance calculation, and principal coordinates analysis (PCoA). Once this is complete, only the results are returned to the frontend, where interactive visualizations are rendered using Plotly.js (see Fig. 2). By keeping all raw abundance data within the Supabase environment, this architecture enforces data minimization and supports compliance with GDPR requirements. This architectural decision directly implements the principles of GDPR Article 25 (Data Protection by Design and by Default). Grounded in Data Protection Impact Assessment (DPIA) logic, the system mechanically enforces data minimization because transmitting highly individual-specific raw abundance tables to client devices presents an unacceptable privacy vulnerability.

For compositional analysis, zero values are handled by adding a user-specified pseudocount to all abundance values prior to CLR transformation; the interface offers pseudocount options ranging from 0.001 to 1. No additional library-size normalization is applied, as the CLR transformation inherently accounts for differences in sequencing depth by referencing each value to the geometric mean of the sample. OTU selection is determined by the user, who specifies the number of OTUs to include in the analysis; these are selected from the dataset in their stored order rather than ranked by abundance. Allowing user-specified pseudocounts at these lower thresholds mitigates the compositional bias introduced when artificially inflating sparse microbiome matrices, ensuring the CLR transformation remains stable without distorting rare taxa.

To verify numerical correctness, the CLR transformation and Aitchison distance calculations were validated against scikit-bio (version 0.7.2). A test dataset was processed using identical pseudocount values in both implementations. The CLR-transformed values and resulting pairwise Aitchison distances were compared, yielding a maximum absolute difference of less than 10^{-15} between the two implementations, consistent with 64-bit floating point precision limits, confirming the accuracy of the edge function calculations.

From a user, perspective the visualization page is accessed through the navigation menu. Upon loading, the page has general statistics about the available datasets including their samples and OTUs. Users can select a dataset and their dimensionality in terms of OTUs and samples. Users can choose between the standard Bray-Curtis dissimilarity or the Aitchison distance, based on CLR transformation (F-101 & F-103). With the latter option they can select a pseudo-count (F-001). At last they can select their color palette (F-015) before generating the visualizations.

The module supports two visualization types for beta diversity analysis. Heatmaps display the pairwise distance matrix, with samples shown on both axes and color intensity representing dissimilarity values (F-005 & F-014). Hovering over the heatmap shows the distance/similarity score (F-017). In addition, PCoA ordination plots display samples as points in two-dimensional space where distance reflects similarity (F-004, partially F-010, NF-004 & NF-101). The axes of the ordination plot report the percentage of variance explained (F-107 & NF-103). At last, samples can be colored according to metadata variables such as treatment group, sex, or age category, to explore correlations (F-016, F-019, NF-002 & NF-102).

Both visualization types incorporate interactive features provided by Plotly.js. Users can hover over elements to view samples and values, zoom in and out to examine specific areas in detail, and change metadata-based color mappings without regenerating the visualization.

3.3 Evaluation

3.3.1 SUS questionnaire

The System Usability Scale (SUS) produced a mean score of 86.8 (SD = 8.9) across all ten participants (see Table 3). Following the framework created by Bangor et al. [5], this score classifies as 'excellent' usability. As this evaluation is formative in nature and participants were recruited from affiliated institutions, the scores should be interpreted as indicative of usability for this proof-of-concept prototype rather than generalizable to broader populations. The quantitative scores

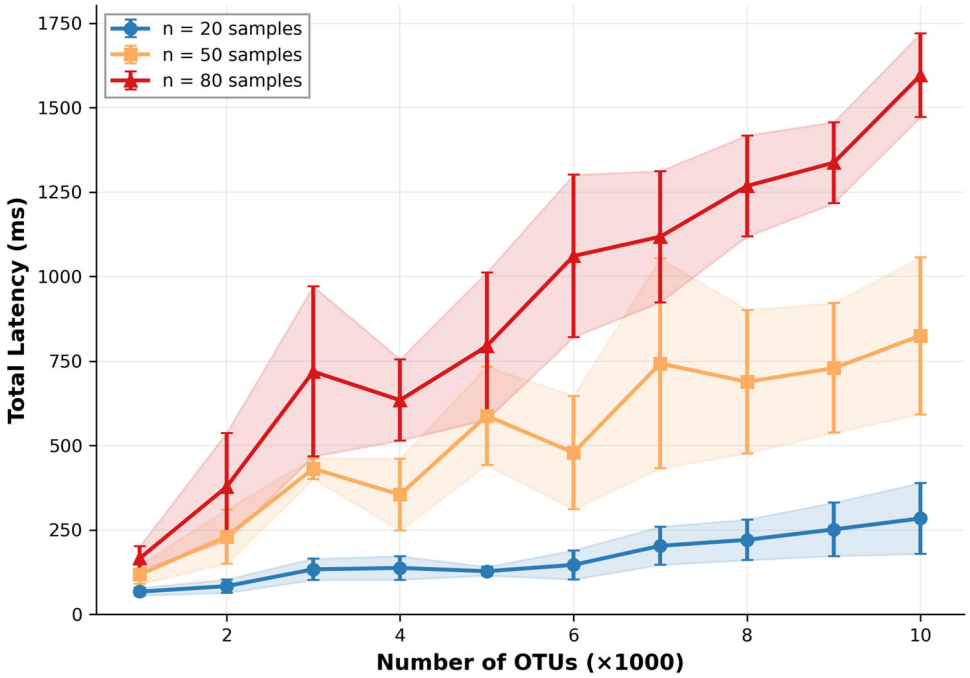

Fig. 1 Edge function latency for the compositionally aware approach across varying dataset dimensions. Total latency (ms) is shown as a function of OTU count for three sample sizes (n = 20, 50, 80). Error bars represent standard deviation across 10 repeated measurements. Latency scales approximately linearly with OTU count, with all configurations remaining below 1,750 ms

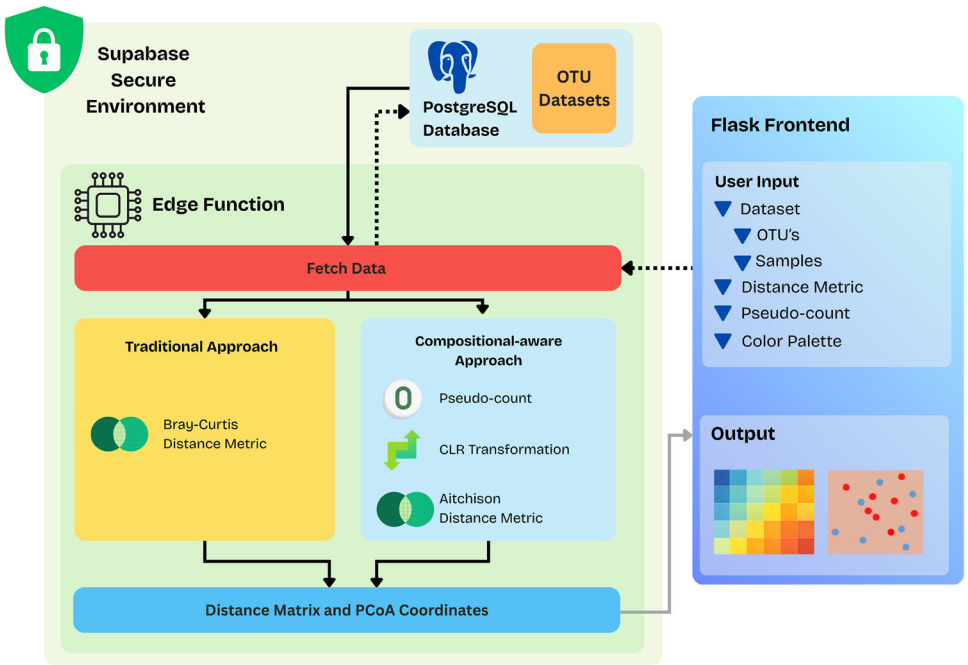

Fig. 2 System architecture for privacy-preserving beta diversity visualization. User-specified parameters are transmitted to a Supabase edge function, which retrieves OTU abundance data from the PostgreSQL database. Within the secure environment, the edge function performs either traditional (Bray-Curtis) or composition-aware (CLR transformation followed by Aitchison distance) calculations, producing a distance matrix and PCoA coordinates. Only the processed results are transmitted to the Flask frontend for rendering, ensuring that raw microbiome data remain within the secure environment

should be interpreted alongside the qualitative feedback from the three interviewed participants.

4 Discussion

4.1 Key findings

The implementation demonstrates that composition-aware beta diversity visualization can be executed entirely within Supabase edge functions, ensuring sensitive data never travels over the network. All core computational steps, including the CLR transformation (F-101), the Aitchison distance calculation (F-103), the pseudo-count handling (F-001), and the PCoA ordination (F-010), are executed server-side with only the processed results transmitted to the frontend for Plotly.js rendering. This setup resolves friction in microbiome research: the need for interactive data exploration conflicts with GDPR requirements that restrict data movement. By processing data at the edge, researchers can generate

Table 3 System Usability Scale (SUS) Scores

Participant	SUS Score
Participant A*	87.5
Participant B*	82.5
Participant C*	82.5
Participant D	97.5
Participant E	100.0
Participant F	97.5
Participant G	85.0
Participant H	80.0
Participant I	72.5
Participant J	82.5
Mean (SD)	**86.8 (8.9)**

*Also completed semi-structured interview

heatmaps and ordination plots without compromising data. The latency results demonstrate that even at maximum load (10,000 OTUs and 80 samples), total response time remained below two seconds, indicating that edge functions provide sufficient performance for exploratory analysis.

The module achieved a mean SUS score of 86.8 (SD = 8.9, n = 10), placing it in the 'excellent' category [5]. While this formative evaluation provides encouraging evidence of usability, results should be interpreted as preliminary given the recruitment from affiliated institutions. Participants consistently rated the interface as intuitive and self-explanatory, with interaction elements positioned in expected locations. The visualization output successfully communicates important information, such as including variance explained, distance metric used, and pseudo-count value, at the point of interpretation, confirming implementation of requirements F-107 and NF-103. This transparency aligns with Munzner's expressiveness principle, which requires visualizations to display all relevant information without misleading the viewer [34].

Participants understanding of a composition-aware method varied greatly, revealing a gap that interface design cannot resolve. Users with a statistical background recognized different distance metric approaches, while others required explanation of the metrics. This suggests that effective composition-aware visualization tools require documentation or guided decision support at the point of method selection.

Despite positive usability scores, participants indicated the module would not replace their specialized workflows. Researchers would like to inspect sample-level distributions before proceeding to beta diversity. Additionally, the current OTU selection mechanism is criticized, and the appropriateness of Aitchison distance for certain microbiome applications was questioned. These findings indicate that standalone beta diversity visualization is insufficient for comprehensive exploratory analysis and establish clear priorities for future development.

4.2 Limitations

This research has several methodological limitations. The evaluation was conducted as a formative, proof-of-concept study. Although the sample size was ten participants for the SUS questionnaire, all participants were recruited from affiliated institutions, which may limit generalizability. According to the Design Science Research approach, artifact development should follow an iterative cycle in which the artifact is evaluated multiple times using stakeholder feedback [39]. This iterative evaluation could not be fully implemented due to time constraints, as it was impractical to involve participants continuously throughout the development of the module. One of the evaluations was conducted remotely via Zoom. The participant was able to use the visualization module through remote control. This setup may have influenced the participant's perceived usability of the system. The other two evaluations were conducted in Dutch, as it was the participant's native language. During the translation of these evaluation, certain nuances may have been misinterpreted, which could have influenced the evaluation results. The created module is unable to encompass the entire landscape of visualization tools. The implementation of additional visualizations is time-consuming, and therefore only beta diversity visualizations are integrated. Furthermore, specific design choices were criticized by participants. The OTU selection parameter was described as "counterintuitive" by Participant B. In addition, the use of Aitchison distance was questioned by the same participant, indicating uncertainty regarding its appropriateness in this context. Finally, all participants were sourced from affiliated institutions. This may have introduced bias, as their background and familiarity with the domain could have influenced their feedback and evaluation outcomes.

4.3 Further research

For future research on expanding the visualization module, the following recommendations are proposed. First, as said by all the participants, the module needs additional visualizations to be useful. Adding relative abundance stacked plots and per-sample histograms could make the module more suitable for their workflow. Integrating in-app tooltips and guided decision support to educate users on when to select Aitchison versus Bray-Curtis metrics. Second, some technical improvements can be made, such as removing the OTU selector and revising the Aitchison distance metric. Another improvement is giving more details about the metrics used, give an overall description of each dataset, and being able to export the visualization with its settings. Third,

as this study represents a formative evaluation within a single DSR iteration, future research should validate these preliminary usability findings with a larger, independent user base drawn from diverse institutions and expertise levels. At last, showing that edge functions are useful for calculating and processing data, it paves the way for more research on this topic. Edge functions are underutilized, and their privacy and security benefits are overlooked. More research on the usage and integration of edge functionality is recommended.

5 Conclusion

This research addressed the challenge of enabling composition-aware beta diversity visualization within FAIRDatabase for human microbiome data. Microbiome research is growing in relevance and with strict privacy constraints, data sharing is limited. The FAIRDatabase was developed to address these challenges, yet lacked a visualization module for exploratory analysis. This study extended FAIRDatabase with a composition-aware beta diversity visualization module implemented using edge functions.

To answer the main research question, *"how can a composition-aware beta diversity visualization module be implemented using edge functions to support researchers with effective exploratory analysis"*, this research demonstrated that implementation is feasible and functional. The module provides an intuitive interface, applies appropriate compositional transformations, and displays relevant information to support interpretation. Nevertheless, its effectiveness depends on individual researcher workflows, and the addition of further visualization types would improve its effectiveness. This research contributes a working visualization module that integrates composition-aware beta diversity analysis into FAIRDatabase using edge functions, demonstrating that edge functions can handle computations while preserving data security, offering a module for privacy-preserving analysis in sensitive research domains. The requirements specification and implementation provide a foundation for future development of microbiome data visualization tools. This work represents a first iteration of the visualization module. Future research should expand the available visualizations and the advantages of using edge functions should be further explored.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41060-026-01107-8.

Author Contributions V.S.M. contributed to the conceptualization and supervision of the study. R.D.v.E. was responsible for the methodology, software implementation, formal analysis, and writing - original draft. S.K. contributed to the investigation, visualization (preparation of figures), and writing - review and editing. All authors contributed to writing - review and editing and read and approved the final manuscript.

Data Availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Aitchison, J.: The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B Stat Methodol. **44**(2), 139–160 (1982)
2. Aitchison, J.: Statistical analysis of Compositional Data. Chapman and Hall (1986)
3. Armstrong, G., Rahman, G., Martino, C., McDonald, D., Gonzalez, A., Mishne, G., Knight, R.: Applications and comparison of dimensionality reduction methods for microbiome data. *Frontiers in Bioinformatics*, 2 (2022)
4. Austin, G. I., Korem, T.: Compositional transformations can reasonably introduce phenotype-associated values into sparse features. *mSystems*, 10(5) (2025)
5. Bangor, A., Kortum, P., Miller, J.: Determining what individual sus scores mean: Adding an adjective rating scale. J. Usability Stud. **4**(3), 114–123 (2009)
6. Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., et al.: Microbiome definition re-visited: Old concepts and new challenges. *Microbiome*, 8(1) (2020)
7. Bindels, L. B., Watts, J. E., Theis, K. R., Carrion, V. J., Ossowicki, A., Seifert, J., Oh, J., Shao, Y., Hilty, M., Kumar, P., et al.: A blueprint for contemporary studies of microbiomes. *Microbiome*, 13(1) (2025)
8. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al.: Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. Nat. Biotechnol. **37**(8), 852–857 (2019)
9. Brooke, J.: Sus: A "quick and dirty" usability scale. *Usability Evaluation In Industry*, page 207–212 (1996)
10. Cao, K., Liu, Y., Meng, G., Sun, Q.: An overview on edge computing research. IEEE Access **8**, 85714–85728 (2020)
11. Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R.: Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. **108**(supplement 1), 4516–4522 (2010)

12. Cho, J.-C.: Human microbiome privacy risks associated with summary statistics. *PLOS ONE*, 16(4) (2021)
13. Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., Xia, J.: Microbiomeanalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Research*, 45(W1) (2017)
14. Dorst, M., Zeevenhooven, N., Wilding, R., Mende, D., Brandt, B. W., Zaura, E., Hoekstra, A., Sheraton, V. M. Fair compliant database development for human microbiome data samples. *Frontiers in Cellular and Infection Microbiology*, 14 (2024)
15. Edgar, R.C.: Uparse: Highly accurate otu sequences from microbial amplicon reads. Nat. Methods **10**(10), 996–998 (2013)
16. Estaki, M., Jiang, L., Bokulich, N. A., McDonald, D., González, A., Kosciolek, T., Martino, C., Zhu, Q., Birmingham, A., Vázquez-Baeza, Y., et al.: Qiime 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. *Current Protocols in Bioinformatics*, 70(1) (2020)
17. Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J., and Huttenhower, C.: Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences*, 112(22) (2015)
18. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., Egozcue, J. J.: Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology, 8 (2017)
19. Gloor, G.B., Wu, J.R., Pawlowsky-Glahn, V., Egozcue, J.J.: It's all relative: Analyzing microbiome data as compositions. Ann. Epidemiol. **26**(5), 322–329 (2016)
20. Greenacre, M.: Compositional data analysis. Annual Review of Statistics and Its Application **8**(1), 271–299 (2021)
21. Greenacre, M.J.: Compositional Data Analysis in practice. CRC Press, Taylor and Francis Group (2019)
22. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Q. **28**(1), 75–105 (2004)
23. Huttenhower, C., Finn, R.D., McHardy, A.C.: Challenges and opportunities in sharing microbiome data and analyses. Nat. Microbiol. **8**(11), 1960–1970 (2023)
24. Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J. R., O'Sullivan, C.: The sequence read archive: A decade more of explosive growth. Nucleic Acids Research, 50(D1) (2021)
25. Kelly, B.: The impact of edge computing on real-time data processing. International Journal of Computing and Engineering **5**(5), 44–58 (2024)
26. Kodama, Y., Shumway, M., Leinonen, R.: The sequence read archive: Explosive growth of sequencing data. Nucleic Acids Research, 40(D1) (2011)
27. Lei, Y., Xiao, Y., Li, L., Jiang, C., Zu, C., Li, T., Cao, H.: Impact of tillage practices on soil bacterial diversity and composition under the tobacco-rice rotation in china. J. Microbiol. **55**(5), 349–356 (2017)
28. Lozupone, C., Knight, R.: Unifrac: A new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. **71**(12), 8228–8235 (2005)
29. Lu, Y., Zhou, G., Ewald, J., Pang, Z., Shiri, T., Xia, J.: Microbiomeanalyst 2.0: Comprehensive statistical, functional and integrative analysis of microbiome data. Nucleic Acids Research, 51(W1) (2023)
30. Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., Peddada, S. D.: Analysis of composition of microbiomes: A novel method for studying microbial composition. Microbial Ecology in Health & Disease, 26(0) (2015)
31. McMurdie, P. J., Holmes, S.: Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE, 8(4) (2013)
32. Miranda, E.: Moscow rules: A quantitative exposé. Lecture Notes in Business Information Processing, page 19–34 (2022)
33. Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., Zengler, K., Knight, R.: Establishing microbial composition measurement standards with reference frames. Nature Communications, 10(1) (2019)
34. Munzner, T., Maguire, E.: Visualization analysis and Design. CRC Press, Taylor & Francis Group (2015)
35. Paulson, J.N., Stine, O.C., Bravo, H.C., Pop, M.: Differential abundance analysis for microbial marker-gene surveys. Nat. Methods **10**(12), 1200–1202 (2013)
36. Pawlowsky-Glahn, V., Egozcue, J.J.: Compositional data and their analysis: An introduction. Geological Society, London, Special Publications **264**(1), 1–10 (2006)
37. Pearson, K.: Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, 60(359–367):489–498 (1897)
38. Peeters, J., Thas, O., Shkedy, Z., Kodalci, L., Musisi, C., Owokotomo, O. E., Dyczko, A., Hamad, I., Vangronsveld, J., Kleinewietfeld, M., et al.: Exploring the microbiome analysis and visualization landscape. Frontiers in Bioinformatics, 1 (2021)
39. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. J. Manag. Inf. Syst. **24**(3), 45–77 (2007)
40. Peterson, C.B., Saha, S., Do, K.-A.: Analysis of microbiome data. Annual Review of Statistics and Its Application **11**(1), 483–504 (2024)
41. Pérez-Losada, M., Narayanan, D. B., Kolbe, A. R., Ramos-Tapia, I., Castro-Nallar, E., Crandall, K. A., Domínguez, J.: Comparative analysis of metagenomics and metataxonomics for the characterization of vermicompost microbiomes. Frontiers in Microbiology, 13 (2022)
42. Ramette, A.: Multivariate analyses in microbial ecology. FEMS Microbiol. Ecol. **62**(2), 142–160 (2007)
43. Shanmugam, G., Lee, S. H., Jeon, J.: Ezmap: Easy microbiome analysis platform. BMC Bioinformatics, 22(1) (2021)
44. Supabase Inc Supabase: The open source firebase alternative. Available at: https://supabase.com. Accessed: 13-11-2025 (2025)
45. Vangay, P., Burgin, J., Johnston, A., Beck, K. L., Berrios, D. C., Blumberg, K., Canon, S., Chain, P., Chandonia, J.-M., Christianson, D., et al.: Microbiome metadata standards: Report of the national microbiome data collaborative's workshop and follow-on activities. mSystems, 6(1) (2021)
46. Wiegers, K. E., Beatty, J.: Software requirements 3. Microsoft Press (2013)
47. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific Data, 3(1) (2016)
48. Zakrzewski, M., Proietti, C., Ellis, J.J., Hasan, S., Brion, M.-J., Berger, B., Krause, L.: Calypso: A user-friendly web-server for mining and visualizing microbiome–environment interactions. Bioinformatics **33**(5), 782–783 (2016)