*Bioinformatics*, , 1–4
doi:

# HoloFoodR: a statistical programming framework for holo-omics data integration workflows

**Tuomas Borman**[1,†] **Artur Sannikov**[2,†] **Robert D. Finn**[3] **Morten Tønsberg Limborg**[4] **Alexander B. Rogers**[5] **Varsha Kale**[6] **Kati Hanhineva**[7] **and Leo Lahti**[8]

[1,8]Department of Computing, University of Turku, 20014 Turku, Finland, [2,7]Department of Life Technologies, University of Turku, 20014 Turku, Finland, [3,5,6]Wellcome Genome Campus, EMBL-EBI, CB10 1SA Hinxton, Cambridgeshire, United Kingdom and [4]Center for Evolutionary Hologenomics, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, 1353 Copenhagen, Denmark

[†]= Equal contribution.

## Abstract

**Summary:** Holo-omics is an emerging research area that integrates multi-omic datasets from the host organism and its microbiome to study their interactions. Recently, curated and openly accessible holo-omic databases have been developed. The HoloFood database, for instance, provides nearly 10,000 holo-omic profiles for salmon and chicken under controlled treatments. However, bridging the gap between holo-omic data resources and algorithmic frameworks remains a challenge. Combining the latest advances in statistical programming with curated holo-omic data sets can facilitate the design of open and reproducible research workflows in the emerging field of holo-omics.

**Availability and implementation:** HoloFoodR R/Bioconductor package and the source code are available under the open-source Artistic License 2.0 at the package homepage https://github.com/EBI-Metagenomics/HoloFoodR.

**Contact:** microbiome@utu.fi

**Supplementary information:** Supplementary data is available in the package vignette via the package homepage https://ebi-metagenomics.github.io/HoloFoodR/articles/HoloFoodR.html

**Key words:** holo-omics, multiomics, metagenomics, metabolomics, data integration, bioconductor

## Introduction

The rapid advancement of omics technologies, including (meta)genomics and metabolomics, has been enhanced by breakthroughs in computational methods (Moreno-Indias et al. 2021; Marcos-Zambrano et al. 2023; Santamaria et al. 2024). These developments have enabled holo-omics, rapidly emerging field, which uses an integrative approach to comprehensively collect and analyze omic data from both a host organism and its associated microbiome, collectively referred to as the *holobiont* (Nyholm et al. 2020; Limborg et al. 2018; Odriozola et al. 2024). The holo-omic approach has improved our understanding of complex biological systems, for example, in aquaculture (Limborg et al. 2018; Brealey et al. 2024).

Holo-omic research and the collaborative development of computational methods could benefit from curated, open access data resources. The limited availability of comprehensive multi-omic data resources can impede research progress, constraining computational method development. Open data portals with an accessible Application Programming Interface (API) are crucial in overcoming these challenges. They facilitate research by providing access to data resources and supporting the collaborative development and benchmarking of new data science methods to extract insights from large, curated datasets (Pasolli et al. 2017). Data can be limited in value unless it adheres to the FAIR principles (Findability, Accessibility, Interoperability, Reuse) (Wilkinson et al. 2016). API interfacing and data wrangling require specialized skills, leading to non-transferable, error-prone workflows, not easily reusable by the wider scientific community. Therefore, standardized open-source workflows are needed to search, retrieve and convert data into a suitable format for downstream analyses. To narrow the gap between upstream data retrieval and downstream analysis, we developed the HoloFoodR package, which facilitates seamless programmatic linking of the holo-omic data via the HoloFood data portal API and analysis

 

methods from the R/Bioconductor (Gentleman et al. 2004; Callahan et al. 2016).

## Materials and methods

The HoloFood data portal is an open access portal of curated holo-omic data and analyses, developed by the international HoloFood consortium (Rogers et al. 2025) and hosted by the European Bioinformatics Institute (EMBL-EBI). It centralizes access to heterogeneous data resources including the European Nucleotide Archive for sequence data, MGnify for metagenomic data, and MetaboLights for metabolomic data, and tracks their interrelations via a web portal and free API, covering nearly 10,000 samples from over 2,000 individual chickens and salmon. Biomolecular and physiological measurements were collected at the level of each individual specimen in the project to explore the effects of novel sustainable feeds on physiological processes in farmed animals. In addition, the metadata of each sample is stored in the EMBL-EBI BioSamples service (Courtot et al. 2018).

We developed a conceptual framework to support best practices of statistical programming in holo-omic research, and implemented the work as the HoloFoodR R software library. The package relies on specialized data containers that have been designed to organize the multi-omic data in a structured format. Once the data has been imported to these formats, users can leverage state-of-the-art statistical programming methods for data processing and analysis.

### Integrative data containers

Data containers provide structured, standardized storage for data, proving useful in life science informatics where complex, hierarchical and multi-source data collections are common to describe the studied phenomenon (Drnevich et al. 2024). Custom data containers simplify handling diverse omics data, supporting the design of modular and reproducible workflows.

Two data structures are central to our approach, encompassing the needs to organize data for a single omics type, and subsequently link these across multiple omics. First, the TreeSummarizedExperiment (TreeSE) data container (Huang et al. 2021) standardizes single-omic data, such as metagenome or metabolome derived data sets. It stores experimental data, along with feature and sample metadata, in a structured format. Additionally, it supports the integration of hierarchical information on the features and samples, such as phylogenetic trees. The second container, MultiAssayExperiment (MAE) (Ramos et al. 2017) acts as a logical complement by integrating multiple, heterogeneous omics datasets, seamlessly linking samples across single-omic datasets. Together, these data containers form the basis for an ecosystem of interoperable methods developed by the R/Bioconductor community.

### HoloFoodR workflow

The HoloFoodR package structures data in these single and multi-omic formats, enabling efficient data importing from holo-omic databases for streamlined analysis and integration. The HoloFoodR workflow is summarized in Fig 1. The package uses four functions to query and retrieve data from the HoloFood and MetaboLights databases (Table 1). A key feature is its ability to organize versatile data combinations into the MAE data container. Thus, it transforms complex data from the HoloFood API into a standardized structure for downstream statistical analyses in the Bioconductor data science environment (Ramos et al. 2017).

Furthermore, HoloFoodR leverages the MetaboLights database to retrieve non-targeted metabolomic datasets indicated by project identifiers in the HoloFood data. The data can then be integrated with complementary metagenomic data resources that are available via the MGnify database using the MGnifyR package (Borman et al. 2024a).

**Table 1.** HoloFoodR functions that facilitate data search and retrieval

| Function | Description |
|---|---|
| doQuery() | Query HoloFood database |
| getData() | Retrieve sample data in list or data.frame format |
| getResult() | Retrieve sample data in MultiAssayExperiment format |
| getMetaboLights() | Retrieve non-targeted metabolomic data from MetaboLights |

### Bioconductor methods ecosystem

When data from the HoloFood database is retrieved via the HoloFoodR package, users gain direct access to single- and multi-omic downstream analysis methods available in Bioconductor. It currently offers 2,300 packages, developed by a diverse community of individual contributors for a wide range of fields (see e.g. Gentleman et al. 2004; Callahan et al. 2016; Amezquita et al. 2020; Drnevich et al. 2024). Of particular relevance for studying host-microbiome interactions, Bioconductor includes the specifically designed `mia` framework centering around these containers and supports a range of analysis and visualization methods in microbiome research (Borman et al. 2024b; Borman et al. 2024c; Borman et al. 2024d). Similarly, the metabolomic data can be pre-processed, analyzed and visualized with the `notame` package (Klåvus et al. 2020).

## Results

In this section, we present a practical use case demonstrating the application of HoloFoodR within a holo-omics workflow. The workflow includes the following steps:

1. Fetch and integrate data from the HoloFood and MGnify databases.
2. Filter, clean, and transform data for analysis.
3. Explore and summarize the data.
4. Test associations between fatty acids, time, and treatment.
5. Test associations between microbiome composition, time, and treatment.
6. Characterize the joint variation between the parallel omics measurements.

The full workflow including figures and data refinement is available in the package vignette at `https://ebi-metagenomics.github.io/HoloFoodR/articles`.

### Data import and wrangling

The workflow begins by querying the HoloFood database to retrieve available animals. We chose to gather all available salmon entries, collecting metadata on these animals to explore the available data types and samples for each. These samples were then compiled into MAE data container for further analysis. For metagenomic data retrieval, we leveraged the capabilities of the MGnifyR package to first identify the corresponding sample IDs from the MGnify database and then fetch the metagenomic data as a MAE object. The consistent data format supports seamless merging of the datasets from HoloFood and MGnify. The resulting

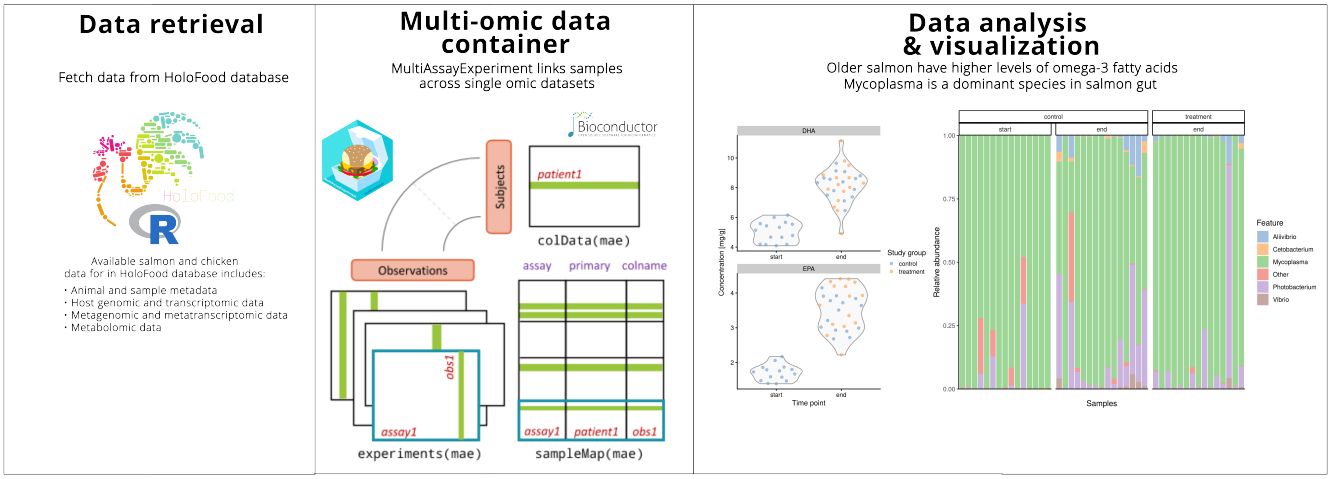


Fig. 1: Data science workflow for holo-omic analysis. Programmatic access to open omic databases is facilitated by the dedicated software that feeds the data to R computational environment. The versatile R/Bioconductor is utilized to visualize microbial genera abundances and to analyze time effect on omega-3 fatty acid concentrations in salmon muscles. The illustration and logo of the multi-assay data container have been taken from the MultiAssayExperiment vignette (Artistic License 2.0).

data container carries various types of omics with associated metadata, and provides the basis for downstream analyses.

By employing the capabilities of the MAE object, we efficiently wrangled the data to focus on salmon from Trial A in the HoloFood database, to study the impact of fermented seaweed as a feed additive. Samples were obtained from animals euthanized during the sampling process, with animals assigned to be sampled either at the beginning or at the end of the trial. Fatty acid concentrations were analyzed from muscle tissue, while metagenomics samples were drawn from the intestine.

The `mia` framework (Borman et al. 2024b; Borman et al. 2024c) provides a user-friendly R interface designed for the downstream analysis, with a particular emphasis on microbiome data. We filtered and agglomerated the data to concentrate on a specific group of microbes and fatty acids. Additionally, various data transformations were applied to prepare the dataset for further analysis. Briefly, we have transformed metagenomic data with centered log ratio method, and fatty acid data with logarithmic transformation. The full details are provided in the package vignette.

### Downstream analysis

After data retrieval and basic exploration, we focused on the effects of the seaweed treatment and aging. We did not observe any impact of treatment on the fatty acid composition, which aligns with the findings of a recent HoloFood study (Rasmussen et al. 2025). However, we did note temporal effects. As the salmon grew, the concentration increased for certain fatty acids, such as docosahexaenoic, eicosapentaenoic, linoleic, oleic, palmitic, and stearic (see Fig 1, data analysis panel, which highlights the first two fatty acids).

Microbial community analysis indicated a dominance of *Mycoplasma*, consistent with previous studies (Zarkasi et al. 2014; Bozzi et al. 2021, see Fig 1, data analysis panel). We also observed an increase in microbial diversity (Shannon index) with time, confirmed by Principal Coordinate Analysis (Bray-Curtis dissimilarity).

To explore the co-abundance of fatty acids and microbial genera, we performed multi-omic factor analysis (Argelaguet et al. 2020), revealing that *Cetobacterium*, *Vibrio*, *Aliivibrio*, and *Photobacterium* covaried with overall fatty acid levels, while *Mycoplasma* showed no such correlation.

## Discussion and conclusions

The integration of multiple omics layers can offer holistic insights into complex living systems. Holo-omics is the new field that emphasizes the interactions between the hosts and their microbiomes. However, such analyses rely on access to curated data collections, computational environments and algorithmic tools. The evolving field of holo-omics can benefit greatly from the collaborative methods development and sharing, thus avoiding the duplication of work (Shetty et al. 2019; Lahti 2018).

To fill the gap between holo-omic data and downstream analysis methods, we developed the HoloFoodR package. It provides programmatic access to the HoloFood data portal and links it with complementary data sources analysis methods from the extensive Bioconductor ecosystem. Our work demonstrates a comprehensive open data science strategy for holo-omics research.

The HoloFoodR software library simplifies data retrieval by minimizing external data wrangling adhering to standardized data structures that support efficient use of the multi-omic analyses methods. This standardization replaces lengthy *ad-hoc* code with concise, and tested open-source solutions. We illustrate the approach with an end-to-end workflow that covers typical steps of a holo-omic data analysis and emphasizes the use of multi-assay data containers to enhance access to methods in the Bioconductor ecosystem.

In addition to method development and analysis, we see potential in using HoloFood data in teaching multi-omics techniques. HoloFoodR offers simplified access to real-world omic data,

enabling more proficient users "learn by doing". The advantage of such data over "toy" data lies in its increased complexity, which provides an excellent foundation for instructors to teach students advanced data cleaning and analysis techniques (Drnevich et al. 2024).

The HoloFood data portal currently provides the interconnectivity between the samples and omics datasets, based on BioSamples identifiers. This linkage provides a template for future multiomics datasets, and as the number of datasets increases, the HoloFoodR package could be generalised so that the concepts can be readily applied to other datasets.

Despite the benefits, there are also limitations. First, raw spectral metabolite data from the MetaboLights database requires extensive preprocessing, often with external tools (Klåvus et al. 2020). Whereas the proposed custom data structures can readily support downstream analyses, their construction and use require sufficient R programming skills.

Thus, the methods and open data science strategy that we have suggested can serve as a template for conducting multi-omic analyses. The HoloFoodR package can be adapted for other data resources relevant to holo-omic research, where the growing Bioconductor ecosystem offers an expanding compilation of data analytical tools.

## Acknowledgements

We are grateful to HoloFood consortium (https://www.holofood.eu) for providing the curated database and API.

## Competing interests

No competing interest is declared.

## Funding

This work was supported by the European Commission in the framework of the Horizon2020 Project FindingPheno [GA 952914] and HoloFood [GA 817729] . A.S. and K.H. were supported by Jane and Aatos Erkko Foundation and the Research Council of Finland [grant numbers 321716, 334814]. M.T.L. was supported by the Danish National Research Foundation [grant DNRF143]. L.L. was supported by Research Council of Finland [grant number 330887].

## References

Amezquita, Robert A. et al. Orchestrating single-cell analysis with Bioconductor. *Nature Methods* 2020;**17**:137–145.

Argelaguet, Ricard et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* 2020;**21**:1–17.

Borman, Tuomas et al. (2024a) *MGnifyR: R interface to EBI MGnify metagenomics resource*. manual. URL: https://github.com/EBI-metagenomics/MGnifyR.

Borman, Tuomas et al. (2024b) *Mia: Microbiome analysis*. manual. URL: https://github.com/microbiome/mia.

Borman, Tuomas et al. (2024c) *miaViz: Microbiome analysis plotting and visualization*. manual.

Borman, Tuomas et al. (2024d) *Orchestrating microbiome analysis*. URL: https://microbiome.github.io/OMA/docs/devel/.

Bozzi, Davide et al. Salmon gut microbiota correlates with disease infection status: potential for monitoring health in farmed animals. *Animal Microbiome* 2021;**3**:30.

Brealey, Jaelle C. et al. Host–gut microbiota interactions shape parasite infections in farmed Atlantic salmon. *mSystems* 2024;**9**:e01043–23.

Callahan, Ben J. et al. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. 2016.

Courtot, Mélanie et al. BioSamples database: an updated sample metadata hub. *Nucleic Acids Research* 2018;**47**:D1172–D1178.

Drnevich, Jenny et al. (Oct. 2, 2024) *Learning and Teaching Biological Data Science in the Bioconductor Community*. DOI: 10.48550/arXiv.2410.01351. arXiv: 2410.01351. URL: http://arxiv.org/abs/2410.01351 (visited on 11/14/2024) pre-published.

Gentleman, Robert C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 2004;**5**:1–16.

Huang, Ruizhu et al. TreeSummarizedExperiment: a S4 class for data with hierarchical structure. *F1000Research* 2021;**9**:1246.

Klåvus, Anton et al. "Notame": Workflow for Non-Targeted LC–MS Metabolic Profiling. *Metabolites* 2020;**10**.

Lahti, Leo (2018) "Open Data Science". In: *Advances in Intelligent Data Analysis XVII* ed. by Wouter Duivesteijn et al. Cham: Springer International Publishing, pp. 31–39. ISBN: 978-3-030-01768-2. DOI: 10.1007/978-3-030-01768-2_3.

Limborg, Morten T. et al. Applied Hologenomics: Feasibility and Potential in Aquaculture. *Trends in Biotechnology* 2018;**36**:252–264.

Marcos-Zambrano, Laura Judith et al. A toolbox of machine learning software to support microbiome analysis. *Frontiers in Microbiology* 2023;**14**:1250806.

Moreno-Indias, Isabel et al. Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Frontiers in Microbiology* 2021;**12**.

Nyholm, Lasse et al. Holo-Omics: Integrated Host-Microbiota Multi-omics for Basic and Applied Biological Research. *iScience* 2020;**23**:101414.

Odriozola, Iñaki et al. A practical introduction to holo-omics. *Cell Reports Methods* 2024;**4**:100820.

Pasolli, Edoardo et al. Accessible, Curated Metagenomic Data through ExperimentHub. *Nature Methods* 2017;**14**:1023–1024.

Ramos, Marcel et al. Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer Research* 2017;**77**:e39–e42.

Rasmussen, Jacob A. et al. A holo-omics analysis shows how sugar kelp can boost gut health in Atlantic salmon. *Aquaculture* 2025;**597**:741913.

Rogers, Alexander B et al. HoloFood Data Portal: holo-omic datasets for analysing host–microbiota interactions in animal production. *Database* 2025;**2025**:baae112.

Santamaria, Guillem et al. Bioinformatic Analysis of Metabolomic Data: From Raw Spectra to Biological Insight. *BioChem* 2024;**4**:90–114.

Shetty, Sudarshan et al. Microbiome data science. *Journal of Biosciences* 2019;**44**.

Wilkinson, Mark D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016;**3**:160018.

Zarkasi, K. Z. et al. Pyrosequencing-based characterization of gastrointestinal bacteria of Atlantic salmon (Salmo salar L.) within a commercial mariculture system. *Journal of Applied Microbiology* 2014;**117**:18–27.