Molecular Omics

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Mol. Omics*, 2024, **20**, 438

Integrating host and microbiome biology using holo-omics

Carl M. Kobel, [a] Jenny Merkesvik, [b] Idun Maria Tokvam Burgos,[c] Wanxin Lai, [b] Ove Øyås, [a] Phillip B. Pope, [abd] Torgeir R. Hvidsten [b] and Velma T. E. Aho *[a]

Holo-omics is the use of omics data to study a host and its inherent microbiomes – a biological system known as a ''holobiont''. A microbiome that exists in such a space often encounters habitat stability and in return provides metabolic capacities that can benefit their host. Here we present an overview of beneficial host–microbiome systems and propose and discuss several methodological frameworks that can be used to investigate the intricacies of the many as yet undefined host–microbiome interactions that influence holobiont homeostasis. While this is an emerging field, we anticipate that ongoing methodological advancements will enhance the biological resolution that is necessary to improve our understanding of host–microbiome interplay to make meaningful interpretations and biotechnological applications.

Received 1st February 2024,
Accepted 10th June 2024

DOI: 10.1039/d4mo00017j

rsc.li/molomics

Introduction

Overview and potential of holo-omics

In many biological systems and environments, both the host and its resident microbiomes are considered as important contributors to the total function of the overall system.[1] To study the biology of a living system, scientists regularly perform ''omics'' analyses of the various biomacromolecules that constitute a living cell, such as DNA, RNA, proteins, and metabolites (Table 1, Fig. 1).[2] Analysis of these different ''layers'' can be further empowered when performed in combination, an approach referred to as ''multi-omics''. The concept of ''holo-omics'' represents an additional thematic shift whereby instead of focusing on molecular details of an individual organism or an isolated reaction in a specific environment, we can consider the biology of a host–microbiome system as a single unit of action. This makes it possible to understand overarching phenomena in the holobiont.[3] In this context, the objective of

[a] *Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway. E-mail: velma.tea.essi.aho@nmbu.no*
[b] *Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway*
[c] *Faculty of Natural Sciences, Norwegian University of Science and Technology, Trondheim, Norway*
[d] *Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, Queensland, Australia*

Carl M. Kobel

Carl is a bioinformatician and a PhD candidate in the MEMO group at NMBU, Norway. Carl's perspective is that microbiomes are largely undervalued and that we should better understand the minute interactions within them. Carl adopts a big data inspired approach, enjoys tinkering with hardware, and building parallelizable bioinformatics pipelines to gain insights into large microbiome datasets.

Jenny Merkesvik

Jenny is a PhD candidate in the Bioinformatics and Applied Statistics group at NMBU. She is part of 3D'omics, a European Union Horizon 2020 project in which she contributes to increase our understanding of host-microbiome interplay through holo-omics. Her work is motivated by the aim of improving animal and feed production, benefitting both the animals and the growing human population, in a sustainable way.

Table 1 Glossary

Term	Definition
Habitat	A defined ecological niche that provides environmental parameters that supports a set of organisms.
Holo-	From Ancient Greek ὅλος: hólos, ''whole''.
Holo-omics	Research that analyses one or more functional layers of omics data from both host and microbiome. The terms holo-omics and hologenomics might be used interchangeably because most omics layers arise from genomic DNA.
Holobiont	An ecological unit consisting of a host and its resident, interacting micro-organisms.
Host–microbiome interface	Any surface where biological features from either host or microbiome can interact.
Integrative analysis	Overlapping or relating the biological factors between two molecular layers or host–microbiome sources.
Metagenomics	Techniques used to study the collective genomic reads from all organisms in an ecological niche.
Multi-omics	Research covering more than one omics layer representing one or multiple interacting organisms. Examples of the former include human multi-omics with measurements that only reflect human biology; and microbial multi-omics without taking the host into account.
Omics	The study of all biomolecules of a specific type. This review focuses on functional omics data, which can be defined as omics data that change over time and across conditions.
Proteomics	Using a bespoke database which is based on *in silico* translation of the genomic sequences, to match mass spectrometric spectra to measure the abundance of proteins in a sample.
Transcriptomics	Techniques used to study an organism's transcriptome, *i.e.* the sum of all of its RNA transcripts.
Untargeted metabolomics	Using methods such as mass spectrometry (MS) or nuclear magnetic resonance (NMR) to measure the abundance of all the metabolites in a sample.

holo-omics is to study biomacromolecules that constitute biological interactions between a host and its microbiome (Fig. 1).

Acquiring a dataset to study host–microbiome interactions is a matter of applying various omics technologies to measure

Idun Maria Tokvam Burgos

Idun Burgos is a PhD candidate in the Systems Biology group at the Norwegian University of Science and Technology. She holds a master's degree in chemical engineering with a specialization in systems biology from the same university. Her PhD project entails studying bacterial communities through genome-scale metabolic networks, applying methods from bioinformatics, biochemical engineering, and systems biology.

Wanxin Lai

Wanxin Lai is a PhD candidate affiliated with both the Bioinformatics and Applied Statistics group and the Faculty of Chemistry, Biotechnology and Food Science at the Norwegian University of Life Sciences (NMBU). Her research focuses on integrating multi omics data through network-based approaches to uncover underlying molecular mechanisms and improve the prediction of phenotypes of interest.

Ove Øyås

Ove Øyås is a researcher in the Microbial Ecology and Meta-Omics group at the Norwegian University of Life Sciences. In 2019, he obtained a PhD in computational systems biology from ETH Zurich, where he developed a passion for understanding biology through model-based integration of omics data. Currently, Dr Øyås is working on multi-omics data analysis and modeling of the rumen gut microbiome as part of the EU-funded HoloRuminant project. Most of his research involves development of scalable computational methods that make it possible to answer new biological questions.

Velma T. E. Aho

Velma Aho, PhD, has been involved in microbiome research since 2013, starting with amplicon sequencing studies and progressing towards increasingly complex multi- and holo-omic projects, constantly striving for a deeper understanding of the roles of microbiomes in mammalian hosts. After ten years of focus on gut microbiota in Parkinson's disease at the universities of Helsinki and Luxembourg, Dr Aho is currently exploring the microbial community of the cattle rumen as part of the Microbial Ecology and Meta-Omics group at the Norwegian University of Life Sciences.

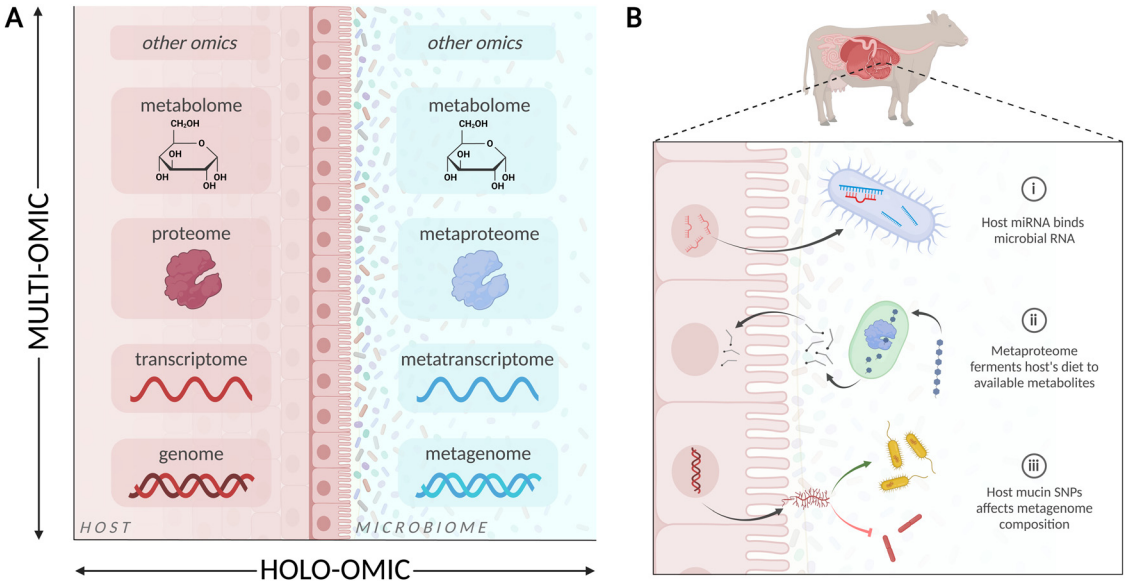

Fig. 1 Holo-omics is a specialised case of multi-omics where biological features are linked across a host–microbiome interface. (A) This interface is idealised along the horizontal axis labelled "holo-omic" as an epithelium with a large surface area where biochemical compounds can be exchanged in both directions. The vertical axis labelled "multi-omic" highlights that interactions can occur on multiple levels in terms of coding sequences and biochemical compounds. (B) Examples of molecular interactions across a host–microbiome interface.[4–6] Created with biorender.com.

the molecular features of both sides of the holobiont. While this data acquisition used to be the limiting step in such analyses, modern molecular biology tools are making this process more efficient and economical. Today's primary technical bottlenecks are (1) overcoming microbial community complexity, which can contain thousands of different genomes compared to their singular defined host, and (2) the computational analysis of holo-omic data so that the biological processes of both the host and its microbiome can be integrated computationally, interpreted, and visualised.[7] For example, performing data integration across the host–microbiome interface requires correlating individual biological features across various omics layers, which often cannot be scaled to the typical size of holo-omic datasets and can also suffer due to insufficient statistical power. To meet this challenge a new family of computational tools is needed: they must be able to cluster biological features into modules and cross-correlate features across the host–microbiome boundary, capturing the signals that represent the hypothesised cooperation between the host and its microbiome.

Symbiotic interactions in host-associated microbiomes are generally defined by the mutualistic, commensalistic or neutral effects shared between each organism, which depend on whether the benefit involved is one-way, two-way or lacking, respectively.[8] Additionally, there is a spectrum of harmful-neutral interactions within the microbiome and between certain microorganisms, and opportunistic pathogens, viruses, and phages might play a role in defining the dynamics of the microbiome.[9] Additional layers of intra-microbiome complexity should also be considered, particularly for the existence of networks of symbioses within a given microbiome which can be characterised in isolation, as with any other microbial environment. What distinguishes holo-omics is that the host variation is integrated together with any intra-microbiome relationships. Subsequently, holo-omics makes it possible to understand the intra-microbiome dynamics where a host-directed interaction is imposed on the microbiome.

For this review, we discuss in detail how actual holo-omic analyses can be performed computationally and present several frameworks to take the typically massive and complex holo-omic datasets and integrate the signal between the host and its microbiome. We consider host–microbiome studies where the host is a multicellular organism like an animal, fungus, or plant that forms a large surface or boundary from which it can interact with the microbiome that typically consists of a community of single-celled microorganisms (bacteria, archaea, eukaryotes) and possibly viruses, with varying degrees of diversity (Table 2). For simplicity, we do not consider parasite interactions in this review but focus on the beneficial interactions in holobionts.

Many known hosts are obligate symbionts, meaning the host is non-viable when the microbiome is absent. One example of an obligate holobiont is lichen, where a fungus and a community of cyanobacteria represent a complete holobiont. The fungus provides physical anchoring and nutrient assimilation whereas the cyanobacteria provide carbohydrates assimilated through photosynthesis. Additionally, these holobionts may house Alphaproteobacteria which work in conjunction to fix nitrogen for the lichen, which may otherwise be nutrient-limited.[30] On the other end of the spectrum of dependency are several types of insects, such as ants and caterpillars, which harbour few or no resident microorganisms that are unlikely to have a large impact on fitness.[31] Mammalian hosts tend to fall between these two extreme examples: they are viable when

Table 2 Selected examples of host–microbiome systems and their characteristics in terms of symbiotic benefit, dependency, species richness, and services exchanged between host and microbiome. These definitions depend on the ecological circumstances in which each host–microbiome system was considered

Holobiont system	Symbiosis	Microbiome richness	Host → microbiome services	Microbiome → host services
Cattle rumen	Mutualistic[8,10]	8500–16 994 prokaryotic species,[11,12] 52 alveolata,[13] 12 fungi[14]	Habitat, substrates[15]	Catabolism of complex plant fibres,[15] anabolism of essential chemicals
Mouse gut	Mutualistic[16]	828–1573 species[17,18]	Habitat, substrates	Catabolism of feed matter, anabolism of essential chemicals[16,19]
Salmon gut	Commensalistic	30–40 species (prokaryotes)[20]	Habitat, substrates	Unknown
Plant root-soil	Mutualistic, commensalistic[21]	2799–271 940 species[22,23]	Energy (sugars, fibres)[24]	Nutrients, nitrogen,[25] stress resistance[21]
Bee gut	Mutualistic	<10 species[26–28]	Habitat, substrates	Modulate social behaviour,[29] catabolism of carbohydrates[28]

raised in a germ-free setting, but experimental results suggest various abnormalities in such animals, ranging from changes in the immune system to altered neurodevelopment and behavior.[32,33]

Host–microbiome orchestration

The holobiont represents an evolutionary shortcut where the host and microbiome partners together orchestrate a metabolic capacity[34,35] that otherwise would have had to develop using horizontal gene transfer and recombination *via* sexual reproduction within the genome of the host organism itself.[36] In the holobiont perspective, the host provides a habitat for its associated microbiomes with defined and stable ecological factors, such as the presence or gradients of substrates and environmental factors like oxygen, temperature, and H^+ concentration.[37] In return, the microbiome provides complex biochemicals that the host otherwise would not have been able to synthesise or assimilate. In this context, host-directed internal environmental factors provide the selective pressure that defines which microorganisms are ultimately present.[38] However, many microorganisms, mainly prokaryotes, utilise promiscuous mechanisms for horizontal gene transfer. This gives them the ability to collect mobile and novel genetic elements from diverse sources such as viruses, and alternative genealogies across domains of life.[39] Mechanisms that enable the rapid evolution of microorganisms facilitate their competitive metabolic potential to assimilate both energy and nutrients from a spectrum of ecological niches. A host that has co-evolved with its microbiome can leverage its microbiome-based metabolic potential flexibility to adapt and thrive in niches that the host would have been unlikely to enter on its own.

The microbiomes of holobionts are per definition not mediated through the somatic genome of the host which means that the microbiota must have its own way of transmitting genetic material to offspring or between individuals in a population. This means that the composition of species present in a microbiome is subject to change over time as new species colonise and take over functions of others.[40] Host–microbiome co-evolution and adaptation is possible when new microbiota become part of the holobiont in a population of hosts, and are inherited vertically to offspring or between individuals in a host population. This can give rise to endemic microbiota species which are exclusively found as part of a holobiont. The microorganisms can adapt to their host and thus diverge from their ancestral population. Hosts and microbiota are able to co-adapt evolutionarily which means that they can each specialise and optimise their function in the holobiont system over generations.[41,42]

Idealised biological frameworks of holo-omic models

Studying holobiont systems using holo-omics generally requires a statistical or mechanistic framework that can capture signals or patterns in the data to infer interacting biomacromolecules or biological features across the host and its microbiome.[43] When analysing holo-omic data, its size and complexity usually means it must first be constrained by dimensionality reduction or compression, or by clustering into modular groups of co-abundant biological features. This is to make the computational analysis tractable and to simplify the interpretation of its function. Therefore, it is necessary that the methodological framework chosen to perform this data constraining is able to capture the hypothesised interaction between the host and microbiome.

Most frameworks are statistical in the sense that they test whether there are significant differences between treatments or co-appearing groups, but suitable mechanistic models are increasingly available and used for data integration as well.[43] To integrate omics data, these mechanistic models should ideally account for the dynamics of all relevant genome-scale networks in the holobiont system, but scaling to systems of this size entails major computational challenges for dynamic models in particular.[44] Because of this, mechanistic omics integration studies have mainly used genome-scale metabolic models (GEMs), which capture the steady-state flows of metabolites through an organism's network of biochemical reactions[45] and are available for a range of hosts and microorganisms.[46] By linking metabolic flows to interactions between host and microbiome, GEMs integrated with holo-omics can allow mechanistic investigation of holobiont systems. Dynamic modelling

of genome-scale interaction networks is also becoming feasible thanks to algorithmic and computational advances,[47] but most of the methods that we will discuss here take a statistical approach where they compare and compute significance between groups.

Examples of recent publications with a holo-omics approach

Since the rise of modern molecular biology tools that have facilitated holo-omic analyses, the number of publications focusing on host–microbiome interactions has been growing. For the purposes of this review, we are particularly interested in studies that include an integrative analysis of two or more omic datasets and discuss both the host and its associated microbiome.

Recent holo-omic research articles provide examples of the different types of questions that can be approached from a holo-omics point of view, ranging from experiments with model organisms to comparative evolutionary studies. In the classic experimental end of the spectrum, two studies used a mouse model to address two ''epidemics'' faced by human medicine: opioid overuse[48] and obesity.[49] Both studies included host transcriptomics, microbial shotgun metagenomics, and untargeted metabolomics, the latter capturing a mix of molecules produced by the host and the microbiome. Their results suggested that the tested medications – morphine in the opioid study, the antidiabetic drug empagliflozin in the obesity study – had effects on the host and microbial layers.[48,49] Both studies further confirmed that there are correlations between different omic layers, offering the simplest kind of evidence for host–microbiome interactions. The opioid study also tested this experimentally by showing that morphine-induced changes in host gene expression vary depending on the presence of a microbiome.[48]

In an example closer to traditional ecology, a study focusing on the gut of the termite *Labiotermes labralis* used metagenomics, metatranscriptomics, and host transcriptomics data to demonstrate that the host and the microbiome provide complementary sets of carbohydrate-active enzymes, enabling the holobiont to degrade a wide range of soil polysaccharides.[50] Finally, a study taking a holo-omics approach to evolution compared several ant- and termite-eating mammals, with findings that supported convergent evolution not only in host genomes, but also in microbiomes.[51] Specifically, the gut metagenomes of these mammals were enriched in enzymes that are necessary for subsisting on an insectivorous diet, such as chitinases and trehalases, compared to mammals with other types of diets.

While the existing publications showcase the exciting opportunities offered by holo-omic research, many of them include only one omics layer for each side of the holobiont. Comprehensive, multi-layered integrative studies remain rare, partly due to financial limitations, but also to the challenges presented by bioinformatic and statistical analyses.

State of the art in integrative models

Considerations for holo-omics tools

Although the cost of generating omics data has come down considerably in recent years, it is still a major undertaking to run controlled animal experiments to obtain matching samples from hosts and microbiomes. As a consequence, holo-omic studies typically tend to have small sample sizes. At the same time, the number of measured biological features (genes, proteins, metabolites) may reach millions, considering that complex microbial communities contain hundreds of species.

Let us consider a hypothetical holo-omic study, where we have measured the host transcriptome of the liver in 100 cows (n = 100) and the meta-transcriptome of the rumen content in those same individuals (p = 20 000 host genes + average 3000 microbial genes × 200 microbial species = 620 000 features). Let us further assume that the experiment is set up to measure methane emission, and that half of the cows were given a methane-inhibiting feed additive (treatment) that indeed reduced emissions. This dataset would pose a massive challenge for data analysis, and not primarily because it would require considerable computational resources to assemble and annotate Metagenome Assembled Genomes (MAGs) and estimate expression (read mapping). The main challenge is related to the large number of features compared to samples. Naively one would think that this dataset could be analysed using multivariate- or machine learning-based prediction methods, where the predictive model could be queried for features or combination of features that contributed significantly to the prediction; ''IF gene G on MAG5 is up AND host gene H is down THEN low methane''. However, with this many features there will be an enormous number of feature combinations that could separate low and high emitting cows, and with only 100 examples (cows) to constrain them, we would never be able to discern real biological feature-combinations from spurious ones (Fig. 2). This phenomenon is referred to as overfitting and is a consequence of the curse of dimensionality: the number of

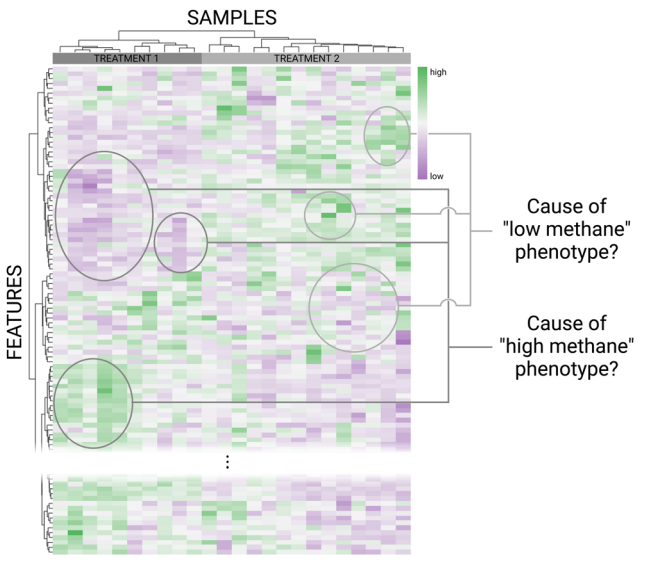

Fig. 2 Illustrating a common problem in multi-omics and holo-omics where a low number of samples with a high number of features are linked into a low number of traits (methane). The underlying data is arbitrary and represents a single omics layer. Created with biorender.com.

examples (cows) needed to identify the biologically meaningful features grows exponentially with the number of features.

Methods that divide the aforementioned examples into training and test sets, such as cross validation, would be able to tell us that we are overfitting, but will not be able to solve the problem. Even testing one feature at a time is problematic, since multiple hypothesis testing would severely limit the statistical power and thus only identify features with very large and consistent differences (*i.e.* large effect sizes) between the two treatments. Luckily, omics features are by no means independent and can be grouped into modules of co-abundant genes, proteins, or metabolites, for instance by correlation. This and other so-called dimensionality reduction approaches typically result in a few dozen distinct modules that can be used as our new features to reveal connections to methane emission and also to hypothesise putative interactions between host and microbiome. A note of caution here is that methods for module finding that rely on computing a distance matrix would require extreme amounts of memory. An approach used for instance by weighted gene co-expression network analysis (WGCNA, a method discussed later in this review) is to first group the data into ''blocks'' using k-means clustering, find modules in each group, and then combine similar modules at the end.

Integrating several omics datasets for a multi-omics approach can help us hone in on biologically meaningful patterns, if done carefully. Assuming that we added metabolomics data to the aforementioned cow example; simply concatenating the transcriptomics and metabolomics table would leave us with even more features (number of genes + number of metabolites). Instead, one could first identify genes and metabolites that are differentially abundant between ''low'' and ''high'' methane-emitting cows, and then select pathways that are enriched in both differential genes and differential metabolites. Such consensus integration methods use information about multiple types of molecules to constrain the number of possible biological interpretations.

Although there are strong functional interdependencies between rumen microbes converting feed into fatty acids and the host animal metabolizing fatty acids in the liver to produce energy, there are also clear physical boundaries separating these features, meaning that we should consider omic data origins in our holo-omic analysis design. In the case of pathway analysis, for example, one needs to consider that a pathway operates within the confines of a cell of a single organism. More generally, most integration methods are designed for a single species, and thus cannot be applied directly in a holo-omics setting. Any pattern discovered in omics data with the aim of describing host-microbiota interactions must include biomacromolecules originating from both sides of the holobiont boundary. This might be accomplished by first applying a standard (multi-)omics analysis method and then filtering the results afterwards, *e.g.* selecting modules containing genes from both the host and the microbiota. However, integrating the host-microbiota constraint as an integral part of the data analysis method could drastically reduce the search space, help deal with the curse of dimensionality and force results to include features from the host that might otherwise drown in the sea of microbial features. The methods described below are selected because we find them especially promising for solving challenges related specifically to holo-omics data sets.

Existing methodological frameworks and tools

Dimensionality reduction. The genetic repertoire of the host and its microbiomes captured by holo-omic data introduces complexities such as data sparsity, sampling variation, ecological differences, and host-specific genetic makeup. Furthermore, distinguishing between free-living and host-associated entities adds another layer of complexity. Since the number of biological features always surpasses the number of observations in holo-omic studies, dimensionality reduction is crucial to create human-interpretable visualisations to explore hidden structures and patterns, and prevent model overfitting.[52] Supervised dimensionality reduction – such as partial least squares discriminant analysis – relies on class labels or response variables to guide the dimensionality reduction process. However, such methods struggle when sample sizes are much smaller than the number of features. On the other hand, unsupervised dimensionality reduction like including matrix factorization and neighbour graph methods, allow discovery of structures in the data without relying on class labels or response variables.[52] Methods that find a few dimensions that are likely to be intrinsic come in two flavours; methods that identify a subset of relevant original features (feature selection), and methods that create new features by combining the original features (feature extraction). Feature extraction methods such as principal component analysis (PCA) (Fig. 3) and single value decomposition utilise variation preservation techniques to extract new features – so-called principal components – that are linear combinations of the original features. Principal components are commonly used for visualising clustering patterns and interpreting sample separation.[53]

Canonical correlation analysis (CCA) is a statistical technique akin to PCA in terms of finding a linear transformation of the original variables that consists of orthogonal vectors.[54] The objective of CCA is to summarise the linear relationship between two sets of variables by identifying linear combinations – called canonical variables – that maximise correlations based on pairs of loading vectors. Although CCA is not primarily designed for dimensionality reduction, it plays a crucial role in comprehending multivariate relationships by revealing the directions in which two sets of variables are most interdependent. Several extensions of CCA further enhance its applicability: (i) multiset CCAs analyse maximal correlations across multiple sets of omics data; (ii) sparse CCAs identify a subset of variables most relevant to the canonical variables by introducing sparsity constraints; (iii) regularised CCAs incorporate regularisation which is particularly beneficial when dealing with high-dimensional data or when variables are not well-captured by linear transformations; and (iv) partial least squares CCAs which focus on predicting one set of variables using another, thus combining aspects of partial least squares

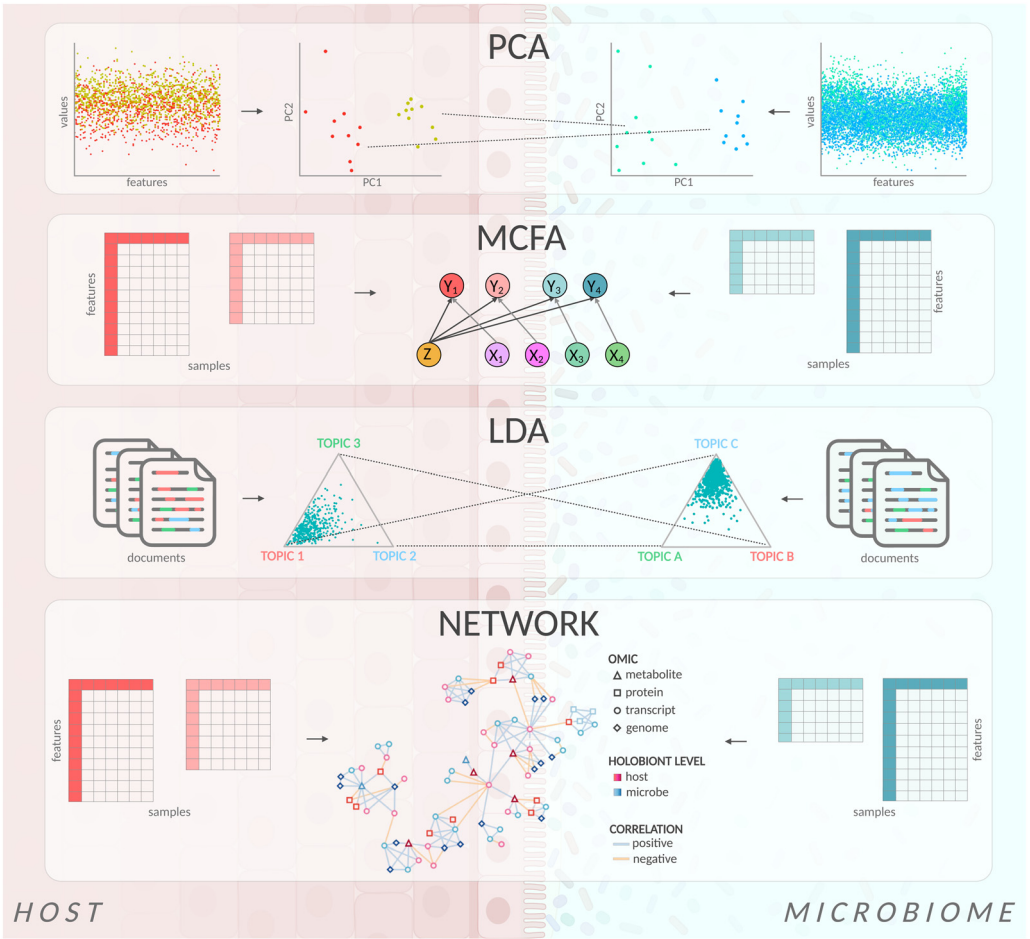

Fig. 3 Figurative summary of the methods discussed in this review. All can reduce inputs with many features to a smaller number of components in order to simplify interpretation of the underlying biological phenomena. PCA: Principal component analysis; MCFA: Multiset correlation and factor analysis; LDA: Latent dirichlet allocation. Created with biorender.com.

regression with CCA.[55] These extensions cater to diverse scenarios, offering flexibility to address specific challenges in multivariate analysis and canonical correlation.

Principal coordinates analysis (PCoA) is a linear transformation method similar to PCA which incorporates multidimensional scaling, creating dissimilarity matrices to visualise sample relationships.[56] Unlike PCA, PCoA is not limited to Euclidean measures and has been shown to be useful for comparing beta-diversity in microbial contexts. Non-metric multidimensional scaling (nMDS) is popular for amplicon/shotgun sequencing data, offering a rank-based approach that handles non-linear relationships and outliers effectively, albeit with potential distortions in global structures.[57–60] Non-linear methods like *t*-distributed stochastic neighbour embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP) belong to the second type of dimensionality reduction, known as neighbour graph algorithms.[59–62] These methods emphasise preserving local structures, relying on graph layout algorithms to create probabilistic weighted graphs representing relationships between high-dimensional data points. UMAP and *t*-SNE differ primarily in their theoretical foundation for balancing the local and global structures.[53] While *t*-SNE results can vary between runs due to its stochastic nature and sensitivity to initialisation, UMAP, although also stochastic, tends to demonstrate more stability across runs. UMAP excels in preserving the global structure of the final projection while still capturing local relationships, it is hence a better choice for prediction tasks.[59,60] Nonetheless, it may struggle to distinguish closely nested clusters. It is crucial to note that all three non-linear methods are sensitive to initialisation, and it is recommended to employ the first two principal components from the linear approach as seeds for initiation. Users should implement these exploratory methods with caution, exploring various hyperparameters, running multiple projections for stability. When choosing a non-linear dimensionality reduction method, careful consideration of data scale, characteristics, and specific research goals is essential.[63]

Matrix factorisation (NMF and MCFA). Aforementioned methods for dimensionality reduction by matrix factorisation – such as PCA – enable compression of large datasets into a smaller feature space, and may thus facilitate identification of important biological factors for the variation in the observed

data. This is particularly relevant for holo-omic studies utilising a matrix factorisation approach, in which we consider complex systems through assembling a variety of data types from both sides of the holobiont, adding to the already prevalent imbalance of few biological samples and high feature counts. Challenges arise when size and heterogeneity of the dataset increases, which calls for adaptations of these matrix factorisation methods when applied in holo-omics.

Non-negative matrix factorisation (NMF)[64] is a method for dimensionality reduction that has been used both in several multi-omic studies and as a basis for additional tools for multi-omic data integration and analysis.[65–69] NMF has the same foundation as PCA, essentially decomposing a large data matrix (D) consisting of feature values (p) across biological replicates (n) into a reduced set of (r) linear expressions. These expressions are represented by two matrices smaller than the original data; one with weights (W, $p \times r$) and one with the reduced feature components (F, $r \times n$) (Fig. 4A).

In contrast to PCA, NMF requires the decomposition matrices to contain non-negative values only. This constraint causes the NMF-derived linear expressions to only consist of addends, thereby preventing cancellations between biological factors with opposing signs. NMF thus reflects the idea of assembling parts – analogous to the omic data layers – into a larger image representing the whole system. Simultaneously, the non-negativity constraint of NMF necessitates the compressed data to be seen as an approximation ($\approx$) of the real data rather than as an equality ($=$).[70] Our objective function for determining the decomposition matrices then becomes to minimise the difference between the real data (D) and the approximation (WF). This iterative approach may yield different solutions based on the initial weight and reduced component matrices, potentially affecting the outcome of the analysis.[71] Hence dimensionality reduction by NMF may be more in line with the analogy of assembling omic datasets to uncover interactions between layers of the complex system, although resulting in an approximated model with a potentially large residual difference caused by the lossy factorisation.

Another approach to holo-omic dataset integration based on matrix factorisation is multiset correlation and factor analysis (MCFA)[72] (Fig. 3 and 4C). While also seeking to compress observed data (D) into matrices for weights (W) and reduced components (F), MCFA effectively divides the model into two parts. One set of decomposition matrices fit the so-called shared space (S), consisting of reduced features with implied importance across all the included omics layers. This shared space is determined through an extension of CCA called probabilistic CCA (pCCA), and it serves the same purpose as the general decomposition seen in NMF. Additional sets of decomposition matrices are then fitted for each individual omics layer through factor analysis, based on the residual between the read data (D) and the modelled shared space. These "private" aspects of the model reflect contributions from factors that are only perceived as important for observations in specific omics layers. The full model then combines the shared and private spaces to approximate the real data, determining the weight and feature matrices through an expectation maximisation algorithm, with the remainder (ψ) being quantified a third addend to complete the expression.

By fitting the observed data to shared and private reduced features separately, the MCFA method may help distinguish between components with implied importance across all levels of the holobiont and those that only appear relevant for a particular omics layer. Additionally, introducing a private model layer for each omic may leave a smaller residual than had the model only covered components relevant for all included data layers. At the time of writing, MCFA has not been applied in a peer-reviewed study since its publication in August 2023, thus its versatility for holo-omic data integration has yet to be demonstrated.

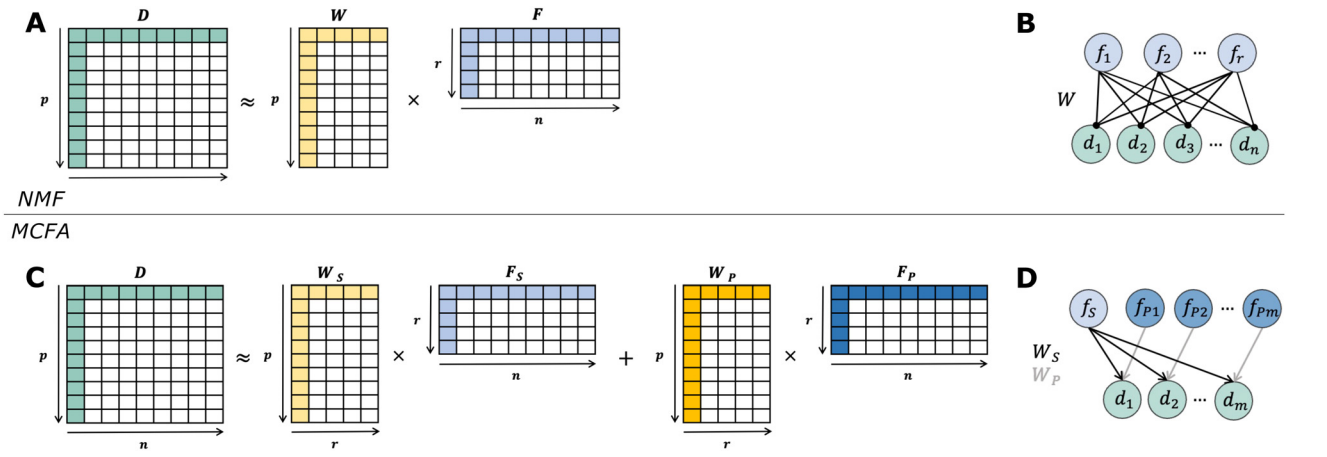

Fig. 4 Comparison of two methods for matrix factorisation; (A) and (B) non-negative matrix factorisation (NMF) and (C) and (D) multiset correlation and factor analysis (MCFA). Both methods reduce a full set of observed data d (columns of D) into linear expressions of reduced features f (columns of F) transformed by multiplication with weights (W). (A) and (C) In contrast to NMF, the MCFA method reduces the dataset into two spaces, either shared between all omics layers (S) or private to each one (P). (B) and (D) All features contribute to approximate the observed data for each shared omics layer, visualised in the same style as Fig. 4 in ref. 66.

Network analysis

Networks are graphs that represent complex relationships between interacting entities within a system.[73] The network is a ubiquitous concept in informatics that can represent many analogous systems like social interactions, flow of information, internet connections, and biological systems like genome-scale metabolic networks (GEMs), genomic co-occurrences, RNA regulation, protein–protein interactions, and metabolics-driven networks.[74] We use correlation networks as an example in this review, as they are suitable for holo-omic studies (Fig. 3). Multi-omics is a more mature concept than holo-omics, hence network methods for the former study type are more developed.[4] We suggest extending these multi-omic tools by integrating data crossing the holo-omic boundary as if it were another omics layer. In general, network analyses handle high-dimensional data well and can provide more interpretable results – compared to other approaches – in the form of node and edge statistics.

WGCNA is a popular framework for investigating associations between biological features within a single omics layer.[75] It calculates an adjacency matrix containing transformed, pairwise correlations between biological features such as genes, proteins, and metabolites. The adjacencies are transformed in order to obtain a scale-free network, in which features can be related to continuous and categorical external data like phenotypic traits or treatment groups. On the basis of these adjacencies, the topological overlap measure can be used together with hierarchical clustering to obtain a set of clusters where each biological feature becomes part of only one of the formulated clusters. In WGCNA terminology, these clusters are referred to as ''modules'' and are represented by their first principal component. This linear combination of biological features is referred to as an eigengene and is idealised to capture the most important variation of the module with limited noise. Since these modules are called without utilising information about treatments or traits, the method can be characterised as unsupervised.

WGCNA can be extended to holo-omic data[76] by relating the modules across the host–microbiome boundary. WGCNA has been applied for both clustering and dimensionality reduction in several multi and holo-omic studies related to both plant[77] and animal biology.[76,78–80] One study concerning the gut microbiome in patients with insulin sensitivity or resistance[79] applied a range of node selection and dimensionality reduction methods on their data, and used WGCNA to find clusters of hydrophilic and lipid metabolites. These were later connected to other omics layers to identify clusters associated with metabolism of the gut microbiome between the groups of patients.

Alternative clustering methods can also be employed for dimensionality reduction. A state-of-the-art example is the Leiden algorithm,[81] which is an optimisation-based form of clustering. The algorithm was used in a study of HIV patients in which they investigated health in relation to the microbiome of the patients. Specifically, they used the Leiden algorithm to detect clusters of microbiome-derived metabolites before integrating these features with other omics layers.[82] Similarly, a study of the SARS-CoV-2 used the Leiden algorithm to detect clusters of metabolites.[80]

Transkingdom network analysis (TkNA)[83] for holo- and multi-omics is a network-based method that detects biological features that differentiate treatment groups. TkNA is designed to handle a binary testing condition, such as ''disease'' and ''control''. The method consists of a comprehensive pipeline containing all the functions needed to transform normalised data into a network that can be readily visualised. TkNA creates a co-variation network and calculates node statistics like node degrees and bipartite betweenness centrality (BiBC). This approach emphasises that hub nodes with high BiBC and degree represent potential modules of the biological network. Additionally, TkNA interfaces with the Infomap[84] and Louvain[85] network clustering algorithms, which can aid in the interpretation of a biological network further.

The size and complexity of networks created from holo-omics datasets make them hard to interpret, hence it is necessary to find ways to categorise and structure the represented data. Clustering nodes and thus reducing the number of visual features to consider can help organise the network. This is exemplified in the aforementioned SARS-CoV-2 study where WGCNA was used to recognise clusters across omics layers. The cross-omic clusters correlating with disease severity revealed a relationship between host serum metabolites and micro-organisms.[80]

In gene set enrichment analysis (GSEA), a gene set usually represents a metabolic pathway that performs a specific biological function. By testing whether there is an enrichment of genes from a specific pathway in a network cluster or module, we can argue that this pathway is captured by the module, thereby drawing further conclusions about its activity by interpreting the module's omics profile and association to other phenotypic metadata. GSEA can be applied on clusters that are defined using any clustering algorithm. An example is a study on the Atlantic salmon[76] where gene enrichment analysis was used to show that certain host RNA genes responded to long chain fatty acids in the feed. A similar method[86] for improving interpretability is network enrichment, in which functional information and network connectivity is integrated. Instead of testing for a significant difference between treatment groups like GSEA, network enrichment quantifies the differential representation among neighbours in the gene network.[87]

A network can be interpreted by statistical concepts that describe crucial properties of the nodes and how they are connected. Degree is simply the number of neighbours of any node. The degree can be expressed relative to the node with the highest number of neighbours, hence degree centrality. Node betweenness describes how many of the pairwise node connections in the network pass through a specific node. If this betweenness measurement is high, the node represents a bottleneck and is indicated to have a potential regulatory effect.[83] The cluster coefficient of a node describes the number of edges between its neighbours in relation to the possible number of edges between these. Coreness considers the neighbourhood of a node as it describes whether a node is part of a

"core" of nodes that are all interconnected with a certain degree (k). Hence, a network can be characterised by the maximum coreness of all nodes. Eigenvector centrality is another network statistic computed for each node in a network. The maximum eigenvalue of the adjacency matrix is computed and is used to normalise the eigenvector, which becomes the eigenvector centralities. Generally, nodes with high eigenvector centrality are essential and interact closely with their respective neighbours.[74] In a study of periodontal disease and response to different treatment, eigenvector centrality was used specifically to find nodes in the network that were connected to other highly connected nodes.[88] This revealed microbial taxa that could be more closely associated with the patients' health status. The same study also looked at the network transitivity – describing the ratio of connected triplets to the number of possible connected triplets – for the networks over different patients and disease states. This statistic is high in the presence of clusters, and the more severe disease cases in the study were associated with lower transitivity. A higher interdependence (*i.e.* transitivity) between microbes was therefore shown to be beneficial for the patient. The severe cases were also more often associated with networks with a high diameter – meaning the shortest path between the most distant nodes – which is expected with low transitivity.

Other tools and frameworks

In addition to the methods introduced above, there are various other multi-omic integration tools that could be useful for holo-omic data analysis. A comprehensive and constantly growing community-maintained list of such tools can be found online in a dedicated Git repository.[89] Aside from a handful of methods aimed at microbiome analyses, this list mainly represents a host perspective. Nevertheless, many of the tools could be used in a host–microbiome context, including the examples highlighted below.

MixOmics[90] is a toolkit that offers both unsupervised and supervised statistical approaches for multi-table datasets, ranging from single omic analysis to complex multi-omics. The supervised method for multi-omics, titled Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO),[91] is based on partial least squares regression/projection to latent structures[92] discriminant analysis (PLS-DA)[93] and sparse generalised canonical correlation analysis (sGCCA),[54] an extension of the CCA method. The sparse version of DIABLO involves using lasso[94] to select those features from each layer that best discriminate between groups of interest. Since DIABLO does not assume any particular distributions from the input data,[91] it is applicable for holo-omic datasets, as long as each layer is normalised in a way that is appropriate to that data type. The limiting factor of this approach is that DIABLO is a supervised method aimed at classification of data into pre-established groups of interest, which makes it less useful for basic, explorative holo-omic studies. Examples where this tool has already been used include a study of the relationships of gut microbiota, dietary fatty acids, and liver gene expression in mice;[95] and the effects of cyanobacterial blooms on the microbiome and metabolome of the medaka fish species.[96]

For studies that do not involve a predefined grouping variable, mixOmics is compatible with mixKernel[97] for multi-omics integration. This explorative, unsupervised approach is based on forming a kernel – a symmetric and positive function that provides pairwise similarities between samples – to represent each layer of data.[97] These can be combined into a meta-kernel by creating one of two alternatives: (i) a consensus kernel, or (ii) a sparse kernel that preserves the topology of the original data. The meta-kernel can then be used in downstream analyses, for example kernel PCA (KPCA)[98] for visualisation of the different layers. Since mixKernel is suited for heterogeneous data, it is also applicable for holo-omics. So far, this method has not been commonly utilised in a host–microbiome context, but it successfully complemented simpler, single-table statistics when selecting plant-beneficial bacterial strains for rice cultivation based on plant growth related measurements.[99]

Another explorative method is mCIA[100] – a multi-table version of co-inertia analysis (CIA or COIA)[101] – which has been tested for selecting rice growth promoting bacteria.[99] CIA resembles sPLS in that it also searches to maximise the covariance between eigenvectors.[100] mCIA has been extended to create sparse mCIA (smCIA) which adds feature selection, improving the interpretability of the results.[102] There is also a further extension, structured sparse mCIA (ssmCIA), which enables incorporating structural information about variables, such as regulatory networks for genes.[102] However, this is less relevant for holo-omic analyses as such pre-existing information is seldomly available.

Compositional omics model-based integration or COMBI[103] is another explorative, unsupervised multi-table method. It is particularly appropriate for host–microbiome analyses since it has been designed to account for compositionality, a feature common to many microbiome measurements such as 16S rRNA gene amplicon data and shotgun metagenomic data.[104] Specifically, compositional data is handled through using the centred log-ratio transform as a link function in the models, while the integrative part of the approach is based on inferring latent variables.[103] This method also offers visualisation of the results as a multiplot showing the features with the largest loadings.

Finally, latent dirichlet allocation (LDA) is a form of unsupervised dimensionality reduction[105] (Fig. 3). It uses a specific terminology as it was originally invented for use in text mining. In a corpus – a set of text documents that represent a spectrum of topics – it allocates each word to a predetermined number of topics so that each word in the total vocabulary belongs to one topic. Each topic is a set of words that, as a whole, revolve around a semantic context. Although the topics are coherent and represent an underlying theme, the title of each topic must be defined manually by interpretation of the listed words in each topic. As a text mining tool, LDA doesn't immediately lend itself useful for biological data inquiries. But, consider substituting a corpus for an omics layer: documents become

biological samples, and genes or compounds become the words. By doing so, the model will be able to capture latent topics defined by biological features that tend to occur together in the same documents (co-abundance), forming topics that represent metabolic functions in the samples. This text-biology analogy means that LDA can be applied for use in biological studies.[106]

Conclusion

As the biological insights of holo-omics are limited by the computational model that picks up host–microbiome interactions, there is a need for better modelling tools. Typically, holo-omic analysis is performed with complex models that use clustering or network analyses coupled with functional enrichment analyses to assign biological functions to interacting groups of biochemical compounds across the host–microbiome boundary. As holo-omics is a specialised case of multi-omics, it is possible to apply multi-omic tools in a holo-omics context. In multi-omics, the omics layers are integrated by correlating clusters of biochemical compounds between layers across the samples. Carried forward, it is possible to integrate the two sides of the holobiont by correlating clusters of biochemical compounds between the host and microbiome sides across the samples.

As this is a new, fast-moving field, there still is no consensus of what is the best way to do science using holo-omics. We hope that this review can generate discussion and new ideas on how to approach the further development of holo-omic methodologies, and we are positive that gold standard methodologies will soon be established.

Author contributions

Conceptualization: CMK, OØ, PBP, TRH, VTEA; funding acquisition: PBP, TRH; supervision: OØ, PBP, TRH, VTEA; visualisation: CMK, JM; writing – original draft: CMK, JM, IMTB, WL, OØ, PBP, TRH, VTEA; writing – review & editing: CMK, JM, IMTB, WL, OØ, PBP, TRH, VTEA.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We gratefully acknowledge the financial support of the European Union's Horizon 2020 research and innovation programme under the grant agreements 101000213-HoloRuminant and 101000309-3D-omics, as well as the Novo Nordisk Foundation under 0054575-SuPAcow.

References

1 J. Roughgarden, S. F. Gilbert, E. Rosenberg, I. Zilber-Rosenberg and E. A. Lloyd, Holobionts as Units of Selection and a Model of Their Population Dynamics and Evolution, *Biol. Theory*, 2018, **13**, 44–65.
2 G. T. Jung, K.-P. Kim and K. Kim, How to interpret and integrate multi-omics data at systems level, *Anim. Cells Syst.*, 2020, **24**, 1–7.
3 L. Xu, G. Pierroz, H. M.-L. Wipf, C. Gao, J. W. Taylor, P. G. Lemaux and D. Coleman-Derr, Holo-omics for deciphering plant-microbiome interactions, *Microbiome*, 2021, **9**, 69.
4 N. Malmuthuge and L. L. Guan, Noncoding RNAs: Regulatory Molecules of Host–Microbiome Crosstalk, *Trends Microbiol.*, 2021, **29**, 713–724.
5 S. L. La Rosa, M. L. Leth, L. Michalak, M. E. Hansen, N. A. Pudlo, R. Glowacki, G. Pereira, C. T. Workman, M. Ø. Arntzen, P. B. Pope, E. C. Martens, M. A. Hachem and B. Westereng, The human gut Firmicute Roseburia intestinalis is a primary degrader of dietary β-mannans, *Nat. Commun.*, 2019, **10**, 905.
6 P. Fan, B. Bian, L. Teng, C. D. Nelson, J. Driver, M. A. Elzo and K. C. Jeong, Host genetic effects upon the early gut microbiota in a bovine model with graduated spectrum of genetic variation, *ISME J.*, 2020, **14**, 302–317.
7 L. Nyholm, A. Koziol, S. Marcos, A. B. Botnen, O. Aizpurua, S. Gopalakrishnan, M. T. Limborg, M. T. P. Gilbert and A. Alberdi, Holo-Omics: Integrated Host-Microbiota Multi-omics for Basic and Applied Biological Research, *iScience*, 2020, **23**, 101414.
8 P. R. Myer, Bovine Genome-Microbiome Interactions: Metagenomic Frontier for the Selection of Efficient Productivity in Cattle Systems, *mSystems*, 2019, **4**(3), DOI: **10.1128/msystems.00103-19**.
9 V. Aggarwala, G. Liang and F. D. Bushman, Viral communities of the human gut: metagenomic analysis of composition and dynamics, *Mobile DNA*, 2017, **8**, 12.
10 I. Mizrahi, in *The Prokaryotes: Prokaryotic Biology and Symbiotic Associations*, ed. E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt and F. Thompson, Springer, Berlin, Heidelberg, 2013, pp. 533–544.
11 M. Kim, M. Morrison and Z. Yu, Status of the phylogenetic diversity census of ruminal microbiomes, *FEMS Microbiol. Ecol.*, 2011, **76**, 49–63.
12 L. Yuan, C. Hensley, H. M. Mahsoub, A. K. Ramesh and P. Zhou, in *Progress in Molecular Biology and Translational Science*, ed. J. Sun, Academic Press, 2020, vol. 171, pp. 15–60.
13 Z. Li, X. Wang, Y. Zhang, Z. Yu, T. Zhang, X. Dai, X. Pan, R. Jing, Y. Yan, Y. Liu, S. Gao, F. Li, Y. Huang, J. Tian, J. Yao, X. Xing, T. Shi, J. Ning, B. Yao, H. Huang and

Y. Jiang, Genomic insights into the phylogeny and biomass-degrading enzymes of rumen ciliates, *ISME J.*, 2022, **16**, 2775–2787.

14 I. V. Grigoriev, R. Nikitin, S. Haridas, A. Kuo, R. Ohm, R. Otillar, R. Riley, A. Salamov, X. Zhao, F. Korzeniewski, T. Smirnova, H. Nordberg, I. Dubchak and I. Shabalov, MycoCosm portal: gearing up for 1000 fungal genomes, *Nucleic Acids Res.*, 2014, **42**, D699–D704.

15 E. Jami, A. Israel, A. Kotser and I. Mizrahi, Exploring the bovine rumen bacterial community from birth to adulthood, *ISME J.*, 2013, **7**, 1069–1079.

16 D. Laukens, B. M. Brinkman, J. Raes, M. De Vos and P. Vandenabeele, Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design, *FEMS Microbiol. Rev.*, 2016, **40**, 117–132.

17 S. Kieser, E. M. Zdobnov and M. Trajkovski, Comprehensive mouse microbiota genome catalog reveals major difference to its human counterpart, *PLoS Comput. Biol.*, 2022, **18**, e1009947.

18 M. Chu and X. Zhang, Bacterial Atlas of Mouse Gut Microbiota, *Cell. Microbiol.*, 2022, **2022**, e5968814.

19 S. P. Rosshart, B. G. Vassallo, D. Angeletti, D. S. Hutchinson, A. P. Morgan, K. Takeda, H. D. Hickman, J. A. McCulloch, J. H. Badger, N. J. Ajami, G. Trinchieri, F. P.-M. de Villena, J. W. Yewdell and B. Rehermann, Wild Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance, *Cell*, 2017, **171**, 1015–1028.e13.

20 A. V.-P. De León, M. Hoetzinger, T. Hensen, S. Gupta, B. Weston, S. M. Johnsen, J. A. Rasmussen, C. G. Clausen, L. Pless, A. R. A. Veríssimo, K. Rudi, L. Snipen, C. R. Karlsen, M. T. Limborg, S. Bertilsson, I. Thiele, T. R. Hvidsten, S. R. Sandve, P. B. Pope and S. L. La Rosa, The Salmon Microbial Genome Atlas enables novel insights into bacteria-host interactions via functional mapping, *BioRxiv*, 2023, DOI: **10.1101/2023.12.10.570985**.

21 B. Bai, W. Liu, X. Qiu, J. Zhang, J. Zhang and Y. Bai, The root microbiome: Community assembly and its contributions to plant fitness, *J. Integr. Plant Biol.*, 2022, **64**, 230–243.

22 H. R. Barajas, S. Martínez-Sánchez, M. F. Romero, C. H. Álvarez, L. Servín-González, M. Peimbert, R. Cruz-Ortega, F. García-Oliva and L. D. Alcaraz, Testing the Two-Step Model of Plant Root Microbiome Acquisition Under Multiple Plant Species and Soil Sources, *Front. Microbiol*, 2020, **11**, DOI: **10.3389/fmicb.2020.542742**.

23 C. R. Fitzpatrick, J. Copeland, P. W. Wang, D. S. Guttman, P. M. Kotanen and M. T. J. Johnson, Assembly and ecological function of the root microbiome across angiosperm plant species, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**, E1157–E1165.

24 A. Pascale, S. Proietti, I. S. Pantelides and I. A. Stringlis, Modulation of the Root Microbiome by Plant Molecules: The Basis for Targeted Disease Suppression and Plant Growth Promotion, *Front. Plant Sci*, 2019, **10**, DOI: **10.3389/fpls.2019.01741**.

25 M. I. A. Cavassim, S. Moeskjær, C. Moslemi, B. Fields, A. Bachmann, B. J. Vilhjálmsson, M. H. Schierup, J. P. W. Young and S. U. Andersen, Symbiosis genes show a unique pattern of introgression and selection within a Rhizobium leguminosarum species complex, *Microb. Genomics*, 2020, **6**(4), DOI: **10.1099/mgen.0.000351**.

26 K. Raymann and N. A. Moran, The role of the gut microbiome in health and disease of adult honey bee workers, *Curr. Opin. Insect. Sci.*, 2018, **26**, 97–104.

27 G. Bonilla-Rosso and P. Engel, Functional roles and metabolic niches in the honey bee gut microbiota, *Curr. Opin. Microbiol.*, 2018, **43**, 69–76.

28 W. K. Kwong and N. A. Moran, Gut microbial communities of social bees, *Nat. Rev. Microbiol.*, 2016, **14**, 374–384.

29 J. Liberti, T. Kay, A. Quinn, L. Kesner, E. T. Frank, A. Cabirol, T. O. Richardson, P. Engel and L. Keller, *The gut microbiota affects the social network of honeybees*, *Nat. Ecol. Evol.*, 2022, **6**, 1471–1479.

30 S. T. Bates, G. W. G. Cropsey, J. G. Caporaso, R. Knight and N. Fierer, Bacterial Communities Associated with the Lichen Symbiosis, *Appl. Environ. Microbiol.*, 2011, **77**, 1309–1314.

31 T. J. Hammer, J. G. Sanders and N. Fierer, Not all animals need a microbiome, *FEMS Microbiol. Lett.*, 2019, **366**, fnz117.

32 P. Luczynski, K.-A. McVey Neufeld, C. S. Oriach, G. Clarke, T. G. Dinan and J. F. Cryan, Growing up in a Bubble: Using Germ-Free Animals to Assess the Influence of the Gut Microbiota on Brain and Behavior, *Int. J. Neuropsychopharmacol.*, 2016, **19**, pyw020.

33 M. Jans and L. Vereecke, A guide to germ-free and gnotobiotic mouse technology to study health and disease, *FEBS J.*, 2024, DOI: **10.1111/febs.17124**.

34 N. A. Moran, H. Ochman and T. J. Hammer, Evolutionary and Ecological Consequences of Gut Microbial Communities, *Annu. Rev. Ecol. Evol. Syst.*, 2019, **50**, 451–475.

35 E. B. V. Arnam, C. R. Currie and J. Clardy, Defense contracts: molecular protection in insect-microbe symbioses, *Chem. Soc. Rev.*, 2018, **47**, 1638–1651.

36 L. Margulis, Serial endosymbiotic theory (SET) and composite individuality, *Microbiol.: Today*, 2004, **31**, 173–174.

37 W. H. Hoover and T. K. Miller, Rumen Digestive Physiology and Microbial Ecology, *Vet. Clin. North Am. Food Anim. Pract.*, 1991, **7**, 311–325.

38 L. G. M. Baas-Becking, *Geobiologie; of inleiding tot de milieukunde*, WP Van Stockum & Zoon NV, 1934.

39 Y. Peng, J. Cai, W. Wang and B. Su, Multiple Inter-Kingdom Horizontal Gene Transfers in the Evolution of the Phosphoenolpyruvate Carboxylase Gene Family, *PLoS One*, 2012, **7**, e51159.

40 I. Mizrahi and E. Jami, A method to the madness, *EMBO Rep.*, 2021, **22**, e52269.

41 E. Rosenberg and I. Zilber-Rosenberg, The hologenome concept of evolution after 10 years, *Microbiome*, 2018, **6**, 78.

42 J. B. Russell and J. L. Rychlik, Factors That Alter Rumen Microbial Ecology, *Science*, 2001, **292**, 1119–1122.

43 E. Noor, S. Cherkaoui and U. Sauer, Biological insights through omics data integration, *Curr. Opin. Syst. Biol.*, 2019, **15**, 39–47.

44 A. J. Lopatkin and J. J. Collins, Predictive biology: modelling, understanding and harnessing microbial complexity, *Nat. Rev. Microbiol.*, 2020, **18**, 507–520.
45 C. Ramon, M. G. Gollub and J. Stelling, Integrating –omics data into genome-scale metabolic network models: principles and challenges, *Essays Biochem.*, 2018, **62**, 563–574.
46 C. Gu, G. B. Kim, W. J. Kim, H. U. Kim and S. Y. Lee, Current status and applications of genome-scale metabolic models, *Genome Biol.*, 2019, **20**, 121.
47 P. Borzou, J. Ghaisari, I. Izadi, Y. Eshraghi and Y. Gheisari, A novel strategy for dynamic modeling of genome-scale interaction networks, *Bioinformatics*, 2023, **39**, btad079.
48 U. Kolli, R. Jalodia, S. Moidunny, P. K. Singh, Y. Ban, J. Tao, G. N. Cantu, E. Valdes, S. Ramakrishnan and S. Roy, Multi-omics analysis revealing the interplay between gut microbiome and the host following opioid use, *Gut Microbes*, 2023, **15**, 2246184.
49 J. Shi, H. Qiu, Q. Xu, Y. Ma, T. Ye, Z. Kuang, N. Qu, C. Kan, N. Hou, F. Han and X. Sun, Integrated multi-omics analyses reveal effects of empagliflozin on intestinal homeostasis in high-fat-diet mice, *iScience*, 2023, **26**, 105816.
50 M. Marynowska, D. Sillam-Dussès, B. Untereiner, D. Klimek, X. Goux, P. Gawron, Y. Roisin, P. Delfosse and M. Calusinska, A holobiont approach towards polysaccharide degradation by the highly compartmentalised gut system of the soil-feeding higher termite Labiotermes labralis, *BMC Genomics*, 2023, **24**, 115.
51 S.-C. Cheng, C.-B. Liu, X.-Q. Yao, J.-Y. Hu, T.-T. Yin, B. K. Lim, W. Chen, G.-D. Wang, C.-L. Zhang, D. M. Irwin, Z.-G. Zhang, Y.-P. Zhang and L. Yu, Hologenomic insights into mammalian adaptations to myrmecophagy, *Natl. Sci. Rev.*, 2023, **10**, nwac174.
52 L. Van Der Maaten, E. O. Postma and H. J. van den Herik, *et al.*, Dimensionality reduction: A comparative review, *J. Mach. Learn. Res.*, 2009, **10**, 13.
53 F. Anowar, S. Sadaoui and B. Selim, Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE), *Comput. Sci. Rev.*, 2021, **40**, 100378.
54 A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Le Cao, J. Grill and V. Frouin, Variable selection for generalized canonical correlation analysis, *Biostatistics*, 2014, **15**, 569–583.
55 X. Zhuang, Z. Yang and D. Cordes, A technical review of canonical correlation analysis for neuroscience applications, *Hum. Brain Mapp.*, 2020, **41**, 3807–3833.
56 J. C. Gower, *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, 2015, pp. 1–7.
57 F. M. Ibarbalz, M. V. Pérez, E. L. M. Figuerola and L. Erijman, The Bias Associated with Amplicon Sequencing Does Not Affect the Quantitative Assessment of Bacterial Community Dynamics, *PLoS One*, 2014, **9**, e99722.
58 J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, 1964, **29**, 1–27.
59 T. T. Cai and R. Ma, Theoretical foundations of t-SNE for visualizing high-dimensional clustered data, *J. Mach. Learn. Res.*, 2022, **23**, 301:13581–301:13634.
60 J. Gauß, Topological and Practical Aspects of Data Separability in Complex High-Dimensional Data.
61 L. McInnes, J. Healy and J. Melville, *arXiv*, 2020, preprint, arXiv:1802.03426, DOI: **10.48550/arXiv.1802.03426**.
62 L. Van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res*, 2008, **9**(86), 2579–2605.
63 M. Rahmatbakhsh, A. Gagarinova and M. Babu, Bioinformatic Analysis of Temporal and Spatial Proteome Alternations During Infections, *Front. Genet*, 2021, **12**, DOI: **10.3389/fgene.2021.667936**.
64 D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, 1999, **401**, 788–791.
65 R. Tappu, J. Haas, D. H. Lehmann, F. Sedaghat-Hamedani, E. Kayvanpour, A. Keller, H. A. Katus, N. Frey and B. Meder, Multi-omics assessment of dilated cardiomyopathy using non-negative matrix factorization, *PLoS One*, 2022, **17**, e0272093.
66 A. R. Kriebel and J. D. Welch, UINMF performs mosaic integration of single-cell multi-omic datasets using non-negative matrix factorization, *Nat. Commun.*, 2022, **13**, 780.
67 Z. Yang and G. Michailidis, A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data, *Bioinformatics*, 2016, **32**, 1–8.
68 S. Mallik, A. Sarkar, S. Nath, U. Maulik, S. Das, S. K. Pati, S. Ghosh and Z. Zhao, 3PNMF-MKL: A non-negative matrix factorization-based multiple kernel learning method for multi-modal data integration and its application to gene signature detection, *Front. Genet*, 2023, DOI: **10.3389/fgene.2023.1095330**.
69 P. Chalise and B. L. Fridley, Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm, *PLoS One*, 2017, **12**, e0176278.
70 A. Akalin, *11.3 Matrix factorization methods for unsupervised multi-omics data integration|Computational Genomics with R.*
71 F. Esposito, A Review on Initialization Methods for Non-negative Matrix Factorization: Towards Omics Data Experiments, *Mathematics*, 2021, **9**, 1006.
72 B. C. Brown, C. Wang, S. Kasela, F. Aguet, D. C. Nachun, K. D. Taylor, R. P. Tracy, P. Durda, Y. Liu, W. C. Johnson, D. Van Den Berg, N. Gupta, S. Gabriel, J. D. Smith, R. Gerzsten, C. Clish, Q. Wong, G. Papanicolau, T. W. Blackwell, J. I. Rotter, S. S. Rich, R. G. Barr, K. G. Ardlie, D. A. Knowles and T. Lappalainen, Multiset correlation and factor analysis enables exploration of multi-omics data, *Cell Genomics*, 2023, **3**, 100359.
73 D. Jiang, C. R. Armour, C. Hu, M. Mei, C. Tian, T. J. Sharpton and Y. Jiang, Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities, *Front. Genet*, 2019, **10**, DOI: **10.3389/fgene.2019.00995**.
74 Z. Liu, A. Ma, E. Mathé, M. Merling, Q. Ma and B. Liu, Network analyses in microbiome based on high-throughput multi-omics data, *Briefings Bioinf.*, 2021, **22**, 1639–1655.

75 P. Langfelder and S. Horvath, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinf.*, 2008, **9**, 559.
76 M. A. Strand, Y. Jin, S. R. Sandve, P. B. Pope and T. R. Hvidsten, Transkingdom network analysis provides insight into host–microbiome interactions in Atlantic salmon, *Comput. Struct. Biotechnol. J.*, 2021, **19**, 1028–1034.
77 J. Xie, Y. Ma, X. Li, J. Wu, F. Martin and D. Zhang, Multi-feature analysis of age-related microbiome structures reveals defense mechanisms of Populus tomentosa trees, *New Phytol.*, 2023, **238**, 1636–1650.
78 B. Czech, Y. Wang, K. Wang, H. Luo, L. Hu and J. Szyda, Host transcriptome and microbiome interactions in Holstein cattle under heat stress condition, *Front. Microbiol*, 2022, **13**, DOI: **10.3389/fmicb.2022.998093**.
79 T. Takeuchi, T. Kubota, Y. Nakanishi, H. Tsugawa, W. Suda, A. T.-J. Kwon, J. Yazaki, K. Ikeda, S. Nemoto, Y. Mochizuki, T. Kitami, K. Yugi, Y. Mizuno, N. Yamamichi, T. Yamazaki, I. Takamoto, N. Kubota, T. Kadowaki, E. Arner, P. Carninci, O. Ohara, M. Arita, M. Hattori, S. Koyasu and H. Ohno, Gut microbial carbohydrate metabolism contributes to insulin resistance, *Nature*, 2023, **621**, 389–395.
80 W. C. Albrich, T. S. Ghosh, S. Ahearn-Ford, F. Mikaeloff, N. Lunjani, B. Forde, N. Suh, G.-R. Kleger, U. Pietsch, M. Frischknecht, C. Garzoni, R. Forlenza, M. Horgan, C. Sadlier, T. R. Negro, J. Pugin, H. Wozniak, A. Cerny, U. Neogi, P. W. O'Toole and L. O'Mahony, A high-risk gut microbiota configuration associates with fatal hyperinflammatory immune and metabolic responses to SARS-CoV-2, *Gut Microbes*, 2022, **14**, 2073131.
81 V. A. Traag, L. Waltman and N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities, *Sci. Rep.*, 2019, **9**, 5233.
82 F. Mikaeloff, M. Gelpi, R. Benfeitas, A. D. Knudsen, B. Vestad, J. Høgh, J. R. Hov, T. Benfield, D. Murray, C. G. Giske, A. Mardinoglu, M. Trøseid, S. D. Nielsen and U. Neogi, Network-based multi-omics integration reveals metabolic at-risk profile within treated HIV-infection, *eLife*, 2023, **12**, e82785.
83 N. K. Newman, M. Macovsky, R. R. Rodrigues, A. M. Bruce, J. W. Pederson, S. S. Patil, J. Padiadpu, A. K. Dzutsev, N. Shulzhenko, G. Trinchieri, K. Brown and A. Morgun, *Nat. Protoc.*, 2024, **19**, DOI: **10.1038/s41596-024-00960-w**.
84 M. Rosvall, D. Axelsson and C. T. Bergstrom, The map equation, *Eur. Phys. J.-Spec. Top.*, 2009, **178**, 13–23.
85 V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech.: Theory Exp.*, 2008, **2008**, P10008.
86 J.-H. Hung, T.-H. Yang, Z. Hu, Z. Weng and C. DeLisi, Gene set enrichment analysis: performance evaluation and usage guidelines, *Briefings Bioinf.*, 2012, **13**, 281–291.
87 A. Alexeyenko, W. Lee, M. Pernemalm, J. Guegan, P. Dessen, V. Lazar, J. Lehtiö and Y. Pawitan, Network enrichment analysis: extension of gene-set enrichment analysis to gene networks, *BMC Bioinf.*, 2012, **13**, 226.
88 L. Sisk-Hackworth, A. Ortiz-Velez, M. B. Reed and S. T. Kelley, Compositional Data Analysis of Periodontal Disease Microbial Communities, *Front. Microbiol*, 2021, **12**, DOI: **10.3389/fmicb.2021.617949**.
89 Github: mikelove/awesome-multi-omics, **https://github.com/mikelove/awesome-multi-omics**, (accessed February 2024).
90 F. Rohart, B. Gautier, A. Singh and K.-A. L. Cao, mixOmics: An R package for 'omics feature selection and multiple data integration, *PLoS Comput. Biol.*, 2017, **13**, e1005752.
91 A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt and K.-A. Lê Cao, DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays, *Bioinformatics*, 2019, **35**, 3055–3062.
92 H. Abdi, Partial least squares regression and projection on latent structure regression (PLS Regression), *WIREs Comput. Stat.*, 2010, **2**, 97–106.
93 K.-A. Lê Cao, S. Boitard and P. Besse, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems, *BMC Bioinf.*, 2011, **12**, 253.
94 R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, *J. R. Stat. Soc., B: Stat. Methodol.*, 1996, **58**, 267–288.
95 M. Schoeler, S. Ellero-Simatos, T. Birkner, J. Mayneris-Perxachs, L. Olsson, H. Brolin, U. Loeber, J. D. Kraft, A. Polizzi, M. Martí-Navas, J. Puig, A. Moschetta, A. Montagner, P. Gourdy, C. Heymes, H. Guillou, V. Tremaroli, J. M. Fernández-Real, S. K. Forslund, R. Burcelin and R. Caesar, The interplay between dietary fatty acids and gut microbiota influences host metabolism and hepatic steatosis, *Nat. Commun.*, 2023, **14**, 5329.
96 A. Gallet, S. Halary, C. Duval, H. Huet, S. Duperron and B. Marie, Disruption of fish gut microbiota composition and holobiont's metabolome during a simulated Microcystis aeruginosa (Cyanobacteria) bloom, *Microbiome*, 2023, **11**, 108.
97 J. Mariette and N. Villa-Vialaneix, Unsupervised multiple kernel learning for heterogeneous data integration, *Bioinformatics*, 2018, **34**, 1009–1015.
98 B. Schölkopf, A. Smola and K.-R. Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Comput.*, 1998, **10**, 1299–1319.
99 M. Truu, S. K. Gopalasubramaniam, G. Muthukrishanan and J. Truu, Application of data integration for rice bacterial strain selection by combining their osmotic stress response and plant growth-promoting traits, *Front. Microbiol*, 2022, **13**, DOI: **10.3389/fmicb.2022.1058772**.
100 C. Meng, B. Kuster, A. C. Culhane and A. M. Gholami, A multivariate approach to the integration of multi-omics datasets, *BMC Bioinf.*, 2014, **15**, 162.
101 S. Dray, D. Chessel and J. Thioulouse, Co-Inertia Analysis and the Linking of Ecological Data Tables, *Ecology*, 2003, **84**, 3078–3089.
102 E. J. Min and Q. Long, Sparse multiple co-Inertia analysis with application to integrative analysis of multi -Omics data, *BMC Bioinf.*, 2020, **21**, 141.
103 S. Hawinkel, L. Bijnens, K.-A. L. Cao and O. Thas, Model-based joint visualization of multiple compositional omics datasets, *NAR: Genomics Bioinf.*, 2020, **2**, lqaa050.

104 G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn and J. J. Egozcue, Microbiome Datasets Are Compositional: And This Is Not Optional, *Front. Microbiol*, 2017, **8**, DOI: 10.3389/fmicb.2017.02224.
105 D. M. Blei, Latent Dirichlet Allocation, *J. Mach. Learn. Res.*, 2003, **3**, 993–1022.
106 C. Tataru, M. Peras, E. Rutherford, K. Dunlap, X. Yin, B. S. Chrisman, T. Z. DeSantis, D. P. Wall, S. Iwai and M. M. David, Topic modeling for multi-omic integration in the human gut microbiome and implications for Autism, *Sci. Rep.*, 2023, **13**, 11353.