



Navigating multi-omic integration methods for human microbiome research

Received: 18 June 2025

Accepted: 16 March 2026

Published online: 21 April 2026

 Check for updates

Efrat Muller^{1,4}, Tal Bamberger^{2,4} & Elhanan Borenstein^{1,2,3}  

Multi-omic studies in human microbiome research hold great potential for advancing our understanding of host–microbiome interactions. However, despite the growing availability of multi-omic datasets, analysing such data remains a major conceptual, analytical and computational challenge. Introduction of new multi-omic integration methods to address these challenges further complicates researchers' efforts to navigate this expanding field. In this Review, we outline the landscape of multi-omic integration methods in the context of human microbiome research. In contrast to previous reviews, we specifically emphasize the different biological questions addressed by various integration approaches, including questions related to interactions between different molecular layers, molecular shifts that occur in disease, subgrouping of patients based on molecular profiles, and identification of biological mechanisms that underlie such associations. Our aim is to provide a timely, convenient and comprehensive resource for the microbiome research community, allowing researchers to identify the multi-omic integration approach that is best suited to their data and objectives.

The human microbiome has become a major focus of biomedical research, with numerous studies uncovering and highlighting its associations with various host health conditions^{1–3}. However, important gaps remain in our understanding of the molecular mechanisms and pathways underlying these associations, hindering our ability to translate findings into effective clinical practices. Multi-omic microbiome studies aim to integrate multiple biological 'omics' to provide a comprehensive, system-level and multilayer perspective on host–microbiome interactions, typically combining metagenomics with additional layers such as metabolomics, transcriptomics, proteomics or host genomics (Box 1). Studies integrating metagenomics with additional omic layers have increased steadily over time, as illustrated by a PubMed-based survey of the literature (Fig. 1 and Supplementary Table 1). Such studies have led to substantial advances in human microbiome research, markedly improving our understanding of microbiome ecology, activity and impact on the host. Combining different omics has been shown, for example, to enhance disease prediction models^{4–7}, improve analyte identification and normalization^{8–12}, and reveal cross-omic

correlations that offer new mechanistic hypotheses^{13–21}. These multi-omic approaches have also enabled researchers to pinpoint functionally active community members within microbiome samples²², assess variability in microbiome activity between individuals that is independent of microbial composition^{23,24}, or identify metabolites whose concentrations are governed by bacterial metabolism²⁵.

Improved technologies and reduced costs have further accelerated the generation of large-scale multi-omic datasets with studies such as the Integrative Human Microbiome Project (iHMP)¹⁷, Lifelines Deep²⁶, PREDICT cohorts²⁷, Swedish CARDiopulmonary bioImage Study²⁸, 500 Human Functional Genomics (500FG)²⁹, TwinsUK³⁰ and MetaCardis³¹, offering exciting opportunities for biological discovery that may not have been feasible with smaller-scale cohorts. While these initiatives highlight the expanding potential for multi-omic research, the analysis of such multimodal datasets remains a major challenge conceptually, analytically and computationally^{32–34}. Data from different omic platforms are highly heterogeneous, varying in format, dimensionality, sparsity (that is, the prevalence of zeros in the data),

¹Blavatnik School of Computer Science and AI, Tel Aviv University, Tel Aviv, Israel. ²Department of Human Genetics and Computational Medicine, Gray Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel. ³Santa Fe Institute, Santa Fe, NM, USA. ⁴These authors contributed equally: Efrat Muller, Tal Bamberger. ✉e-mail: elbo@tauex.tau.ac.il

BOX 1

Common omics in microbiome research and key data characteristics and challenges

16S rRNA sequencing targets the conserved 16S ribosomal RNA (rRNA) gene in prokaryotes, enabling taxonomic identification of bacterial and archaeal taxa within a sample. 16S data typically have limited taxonomic resolution (often genus level) and may include biases related to primer design and the 16S variable region targeted.

Shotgun metagenomics sequencing involves sequencing of all genetic material in a sample, enabling taxonomic and functional profiling as well as de novo assembly of previously uncharacterized genomes. These datasets are typically very large, leading to substantial computational demands.

Metatranscriptomics profiles microbial gene expression and provides insight into active functions. Interpretation is complicated by variation in taxon abundances and gene copy numbers. Datasets are often dominated by rRNA, requiring depletion steps that may introduce bias, and RNA instability can distort expression profiles.

Host transcriptomics measures gene expression in host tissue or cells. Challenges include strong tissue- and cell-type specificity and pleiotropy (where a single gene influences multiple traits).

Metabolomics profiles small molecules that serve as intermediates or end products of metabolic processes. Data can be affected by instrumental drift and variability across instruments and analytical platforms. Measurements are often semi-quantitative, dominated by uncharacterized metabolites and exhibit heteroscedasticity (measurement errors that vary with abundance).

Lipidomics is a subfield of metabolomics and focuses on lipid molecules. It shares many of the same challenges.

Metaproteomics profiles proteins in a sample. Challenges include instrumental drift, heteroscedasticity, limited sensitivity and quantification biases related to ionization, digestion efficiency or peptide detectability. As proteins are inferred from peptides that may

map to multiple proteins, resulting profiles may be incomplete or ambiguous.

Host-genomics analyses host DNA, typically using whole-genome sequencing, whole-exome sequencing, single nucleotide polymorphism (SNP) arrays or other targeted sequencing approaches. Key challenges include linkage disequilibrium (non-random association of genetic variants), pleiotropy and confounding related to population ancestry.

Many omic data types share additional complexities, including high dimensionality (thousands to tens of thousands of features), compositionality (relative rather than absolute abundances), sparsity (most feature measurements are zero) and non-normal distributions requiring tailored transformations or models. Contamination, especially in low-biomass samples, is another concern and may require analysis of negative controls. Sequencing-based omics are also sensitive to sequencing depth, affecting detection limits. Further challenges stem from study design and data processing, including batch effects, dependence on bioinformatic pipelines and reference databases (which may be incomplete or biased), and the computational demands of large datasets.

Other important data types that are often collected in human microbiome studies:

Dietary information estimates the composition and quantity of food intake, typically using self-reported food-frequency questionnaires or food logs. These data are often limited in accuracy and resolution, non-standardized and include mixed data types (categorical, discrete, continuous), complicating analysis and integration with other data modalities.

Electronic health record data include clinical diagnoses, laboratory measurements, medications and other patient information. These datasets are heterogeneous, contain structured and unstructured data (for example, physician notes), are often incomplete and irregularly sampled, and may reflect sampling biases such as 'care-seeking selection bias'.

scale and distribution^{35–37}. Each omic profile generally requires unique processing steps, involving, for example, specific transformations, normalizations or imputation protocols^{35,37–39}. In addition, omic datasets are inherently high dimensional, with considerably more variables than samples, increasing the risk of bias and overfitting^{34,40}. Differences in the number of features in each omic profile may also introduce bias, wherein certain omics may mistakenly appear to be more informative than others⁴¹. Handling these large datasets efficiently requires substantial computational resources^{39,40}, which may not be readily available to all researchers. Finally, while commonly used multi-omic analysis approaches may be effective in identifying general trends, regularities or disease signatures in the data, the interpretation of such findings is often extremely challenging, failing to provide clear mechanistic insights into interactions that occur across different molecular layers and into complex molecular processes³⁹.

A variety of methods have been introduced for multi-omic integration and analysis, aiming to address these challenges. These range from simple statistical tests to machine learning and deep learning models,

network analysis approaches, matrix factorization techniques (which decompose data matrices into lower-dimensional representations) and other methods designed to handle specific omic properties or combinations. Importantly, however, as the landscape of multi-omic integration methods continues to expand, so does the daunting and time-consuming task of navigating it to identify the most appropriate method for a given study objective.

Moreover, in mapping this landscape, it should be noted that methods to integrate multi-omic data can be classified using a few different 'classification axes', including: (1) the research question each method targets; (2) the resolution at which omics are integrated (for example, global omic associations versus associations between specific features); (3) the phase at which integration occurs (for example, early integration where omics are combined before modelling versus late integration where each omic is first modelled and then models' outcomes are combined)⁴², (4) the characteristics of the omic data that can be integrated (for example, supported number of omics, data types, distributions and so on) and finally, (5) the algorithmic

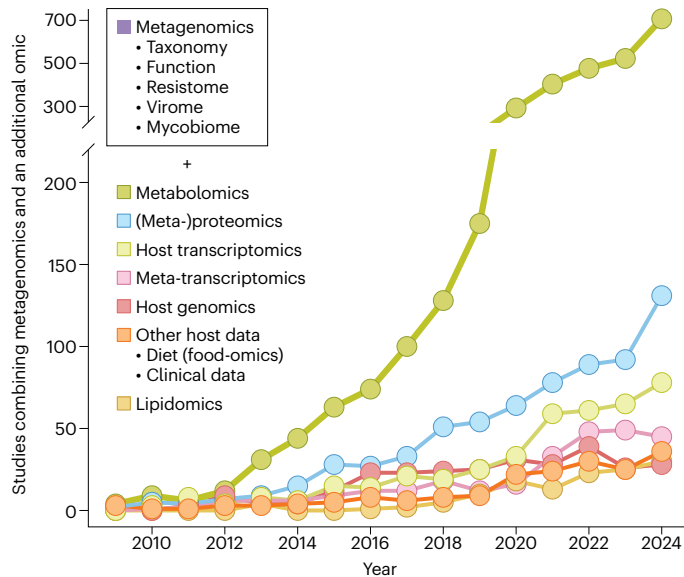


Fig. 1 | Trends in multi-omic human microbiome studies. Number of human microbiome studies per year that combined metagenomics with another specific omic or data type. Numbers are based on PubMed search results, given by the queries detailed in Supplementary Table 1.

approach used for integration (for example, network-based, matrix factorization, probabilistic models, machine learning and so on). Figure 2 (top) illustrates these axes, along with key methodological choices pertaining to each axis. These axes are also largely orthogonal, allowing for nearly any combination of categories across the different axes and further complicating the selection of the best suited method for addressing a specific task.

Comprehensive reviews have already focused on specific omic combinations^{43–45}, particular diseases or environments^{46–50}, or on statistical and algorithmic aspects of data integration^{51–55}, providing some clarity but generally giving less attention to the types of biological insight that different methods can provide. In this Review, we therefore categorized methods primarily on the basis of the research question being targeted by each method (axis 1 in the list above), as this axis provides a useful starting point for selecting approaches that appropriately address the question of interest. We highlight four key research questions central to multi-omic microbiome studies (Fig. 2, bottom): (1) How do different molecular layers interact? (2) How do different molecular layers shift in disease? (3) How are patients stratified on the basis of variation across different molecular layers? and finally, (4) What mechanisms underlie observed statistical associations? Our aim is to provide microbiome researchers, especially those embarking on multi-omic analysis efforts, a clear and comprehensive overview of the current landscape of multi-omic integration methods and to help pinpoint approaches that best align with specific research questions. Supplementary Table 2 further provides a summary of integration methods and their key attributes.

Methods for investigating inter-omic associations

Understanding how different molecular layers, such as the microbiome's genetic composition, metabolome and proteome, interact with one another (for example, exploring how shifts in community composition correlate with changes in the gut metabolome; Fig. 3) requires computational methods to identify statistical associations between different omic layers, either at a global scale or at a more granular level, detecting specific feature-to-feature relationships across omics. A useful application of inter-omic associations is predicting features from one omic based on another, for example, using metagenomic data to

infer metabolite abundances. Methods described in this category help to map the associations between the different omics at various scales and may offer biological hypotheses regarding molecular interactions and processes.

Finding global associations between omics

Finding overall associations between different omics can be challenging as they may have variable scale, dimensionality, distributions and so on. One common approach to assess such 'global' associations relies on sample-to-sample distance matrices derived from each omic, effectively determining whether samples that are proximate according to one omic are also proximate according to the other. These methods allow users to choose specific distance metrics that are the most appropriate for each omic and best account for its unique characteristics. For example, Bray–Curtis or UniFrac⁵⁶ distances could be used for metagenomics-based taxonomic profiles, whereas Euclidean or Mahalanobis distance metrics are commonly used for metabolomics^{57,58}. Conversely, when using non-continuous data, such as count, categorical or binary data, chi-squared or Jaccard metrics can be used.

The omic-specific distance matrices can then be fed into one of several methods to assess global associations between omics (Fig. 3, bottom left). The Mantel test is one such method⁵⁹, evaluating the correlation between distance values from two independent distance matrices using a permutation test to assess the correlation's significance. Mantel tests were used, for example, to confirm a significant association between ancestry and genetic similarity in an Israeli cohort of over 1,000 participants from diverse ancestry backgrounds, and the absence of an association between genetic kinship or ancestry and overall microbiome composition⁶⁰. It was also used to show that functional profiles obtained from metagenomics, metatranscriptomics and metaproteomics were highly correlated with each other in the iHMP study¹⁷. For studies with more than two omics, the partial Mantel test (for comparing two distance matrices while controlling for a third)⁶¹ and multiple regression on distance matrices (MRM)⁶¹ can be used.

While the Mantel test detects linear relationships (or monotonic relationships, when using Spearman correlation), other methods, such as the distance correlation t -test, may be used to detect a more general dependency between two datasets, including nonlinear or non-monotonic relationships⁶². Distance correlation t -tests were used, for example, to map global associations between faecal and plasma metabolomics, faecal and saliva 16S-based taxonomic profiles, and short- and long-term dietary profiles in a cohort of 150 healthy adults¹⁹. This revealed that while the gut and plasma metabolome are significantly associated with both short- and long-term diet profiles, the gut microbiome is associated only with the 'long-term' diet profile.

Procrustes analysis can also be used to assess global associations based on distance matrices⁶³. It uses a pair of ordinations as input, typically calculated using custom distance matrices, and a dimensionality reduction technique such as principal coordinates analysis. Each ordination is then viewed as a 'spatial configuration' of the samples and the method searches for the optimal translating, scaling and rotation scalars so that one configuration best matches the other. A useful advantage of Procrustes analysis is the ability to visually assess the identified superimposition and to identify specific samples with unusual mappings. Procrustes analysis has previously been used to demonstrate a significant overall association between gene expression data from the host colon (using Aitchison's distances) and gut microbiome composition (using Bray–Curtis distances) in patients with colorectal cancer, but not in patients with inflammatory bowel disease (IBD) or irritable bowel syndrome¹³. Although Procrustes analysis is specifically sensitive to outliers and differences in the scales of various features^{64,65}, its performance, similar to that of other distance-based approaches, depends on data transformations, normalizations and distance metrics used⁶⁶.

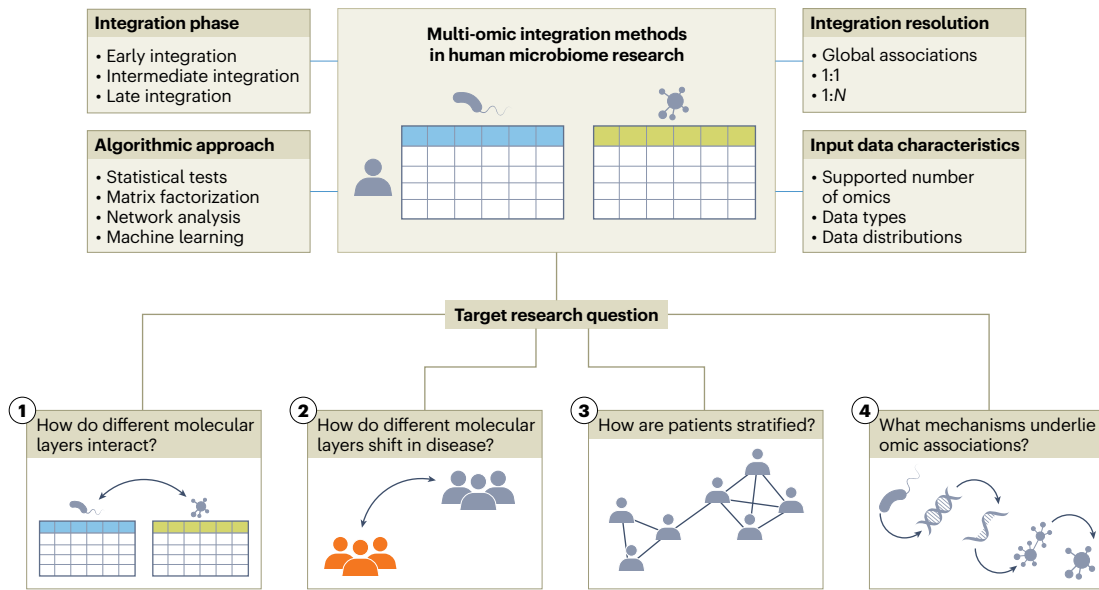


Fig. 2 | Classification schemes for multi-omic integration methods in human microbiome research. Top: several complementary ways to classify multi-omic integration methods, including integration phase, algorithmic approach, integration resolution, input data characteristics and the target research question addressed by the method. Bottom: focus on this last classification scheme, highlighting four common research questions investigated in multi-

omic studies: how different molecular layers interact, how different molecular layers shift in disease or other host phenotypes, how patients can be grouped on the basis of variation across multiple molecular layers and what mechanisms may underlie observed statistical associations between omic layers. Each research question is discussed in the corresponding section of the Review and illustrated in Figs. 3–6.

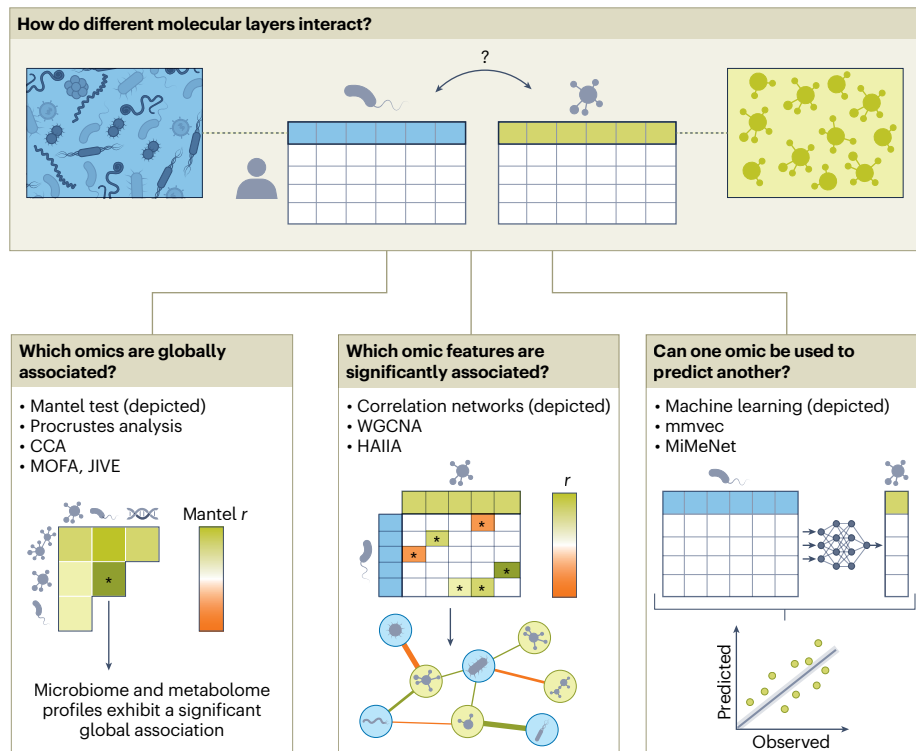


Fig. 3 | Multi-omic integration approaches to determine how different molecular layers interact. Researchers could explore whether different omic profiles are globally associated (bottom left), identify specific features from the various omics that are significantly associated with one another (bottom middle)

or use one omic to predict profiles of another omic (bottom right). Each box lists key methods that can be used to address the specific question, with the method illustrated in that box marked as ‘depicted’.

The aforementioned methods simplify the analysis by focusing on pairwise distances between samples, which also helps in avoiding biases towards larger omics datasets. However, they ignore information about the original features, limiting the interpretability of obtained

results. An alternative approach assumes that high-dimensional data arise from a lower-dimensional space of latent variables. Latent variables may thus represent certain underlying biological processes that cannot be measured directly but that simultaneously affect multiple

measured features across several omic layers. By inferring such latent variables and their relationship across omics, one can identify global associations between omics in the lower-dimensional space or use this simpler space for dimensionality reduction. Methods implementing this approach are collectively referred to as latent variable models, with matrix factorization (MF; the idea of approximating a numeric matrix into a product of lower-dimension matrices) being the most commonly used. Popular MF methods include partial least squares (PLS)⁶⁷, canonical correlation analysis (CCA)⁶⁸, co-inertia analysis (CIA)⁶⁹, joint and individual variation explained (JIVE)⁷⁰, and multi-omics factor analysis+ (MOFA+)⁷¹, each assuming a slightly different underlying model or a different objective function

PLS takes two feature tables as input, one considered to contain 'predictor variables' and the other to contain 'response variables', and seeks to identify latent variables (that is, linear transformations of each feature table separately) that capture the maximal shared covariance⁶⁷. The sparse version of PLS, sPLS, is especially suited for high-dimensional data. It was applied, for example, to data from a study of paediatric allogeneic haematopoietic stem cell transplantation, where operational taxonomic unit abundances were modelled as predictors and clinical variables such as immune markers and immune cell counts as responses⁷². Hierarchical clustering of the sPLS results revealed three clusters, including one defined by rapid natural killer and B cell recovery, high Ruminococcaceae abundance, and an association with mild or no graft-versus-host disease and low mortality. CCA is a similar method but identifies linear combinations of variables in each dataset that maximize their correlation, treating them symmetrically and therefore not assuming a directional relationship⁶⁸. It was used to explore interactions between the gut microbiome and the host transcriptome in infants, uncovering a significant multivariate relationship involving 11 immunity- and defence-related host genes and specific microbiome virulence characteristics⁷³. CIA is another method that generalizes CCA and PLS by maximizing a more general measure known as 'co-inertia' that aims to quantify the degree of co-variability between the latent variables of the two (or more in the generalized version) omics⁶⁹. CIA was used to show remarkable similarity between family members over six different omic profiles, demonstrating conservation of family-specific traits across various omics⁷⁴. Sparse versions of these methods can provide not only a measure of global associations between omics but also information about the most important features contributing to this association. Sparse CCA, for example, uncovered an association between increased *Ruminococcus* and decreased *Veillonella* to higher sphingomyelin and lower lithocholic acid within gut microbiome-metabolome associations of infants predisposed to type 1 diabetes (T1D)⁷⁵. PLS, CCA or CIA models are prone to overfitting and biases when applied to small sample sizes^{76,77}, therefore, validating these models on a set of test data to ensure stability, reliability and generalizability is highly recommended. More recent methods, such as JIVE and MOFA+ have further relaxed assumptions of linearity and do not rely on predefined directional relationships⁷⁸. JIVE decomposes variance in each omic into common ('joint') and omic-specific ('individual') variation, providing an estimation of variation that is explained by these joint components, individual components and residuals⁷⁰. MOFA+ similarly extracts low-dimensional representations from multiple omic profiles to capture both shared and unique omic variation⁷¹. It was used to explore opportunities for personalized healthcare based on continuous digital health data, genomics, proteomics, autoantibodies, metabolomics and gut microbiome profiles⁷⁹. We refer readers to a detailed review on multitable microbiome integration with a focus on matrix factorization⁵³.

Global associations between omics can also be assessed on the basis of omic-specific aggregate variables allowing researchers to use markedly simpler statistical methods such as regression models. For example, researchers often use α -diversity measures to summarize the overall diversity of the microbiome within each sample with a single

variable. A study that analysed the blood metabolome and gut microbiome profiles of 399 individuals from a wellness programme cohort has demonstrated that 40 plasma metabolites, including metabolites of microbial origin, explain approximately 45% of the variance in the gut microbiome's α -diversity⁸⁰.

Identifying associations between specific omic features

Characterizing global relationships between omic layers may indicate overall coordination between processes, but cannot generally pinpoint specific mechanisms or components involved in such coordination. Researchers thus often opt to identify associations between 'specific' features across omics, thereby generating new hypotheses about potential biological interactions or mechanisms (Fig. 3, bottom middle).

A common first step is a simple pairwise correlation analysis (sometimes referred to as 'marginal correlation analysis'), where linear, monotonic or ordinal relationships between features are calculated using simple statistical tests, such as Pearson, Spearman or Kendall rank correlation, respectively. The correlations can be visualized using heat maps or networks that can be clustered to highlight patterns or groups of related features across the dataset. For example, a recent study on site-specific differences between right-sided (RCC) and left-sided (LCC) colon cancer used Spearman correlation to integrate genetic regulation, microbiome profiles and faecal metabolites, highlighting distinct profiles (each linked to different taxa and metabolites) for RCC and LCC²¹. Another recent study used this approach to explore diet patterns (omnivore, vegetarian, vegan) and gut microbiome composition associations in 21,561 individuals. Multiple significant correlations emerged, including an association between red meat consumption in omnivores and an enrichment of microbes such as *Ruminococcus torques*, *Bilophila wadsworthia* and *Alistipes putredinis*, known to be negatively associated with cardio-metabolic health⁸¹. Conversely, microbes enriched in vegans, which were also observed in omnivores with higher plant-based food intake, correlated positively with improved cardio-metabolic markers. Cross-omic pairwise correlations can also be represented as correlation networks, where nodes correspond to features and edges correspond to significant correlations above some user-defined threshold⁵¹. As an example, correlation networks that integrated host genetic traits, clinical tests, metabolomes, proteomes, microbiomes and indicators of physical activity¹⁸ revealed clusters of analytes associated with host physiology and health indicators, enabling the identification of known and novel health biomarkers.

While these approaches employ 'hard thresholding', that is, binarizing pairwise associations as either existing (significant) or not, 'soft thresholding' can also be considered, where associations are weighted by their strength and represented as an edge-weighted network. Methods such as the weighted gene co-expression network analysis (WGCNA) apply this approach to identify modules (clusters) of highly connected nodes within the correlation graph⁸². It was employed to explore how the gut microbiome affects COVID-19 outcomes by identifying modules associated with disease severity within a network constructed from cytokines, metabolites and microbiome features⁸³. WGCNA can also be employed to identify modules within each omic before omic integration, followed by an analysis of associations between the modules across omics. A recent study used WGCNA to analyse the Cancer Genome Atlas (TCGA) tumour microbial and mRNA data, identifying 182 microbiome and 570 mRNA modules⁸⁴. Correlation analysis between the microbial and mRNA modules revealed key microbiome members linked to tumour immune modulation and prognosis.

The extreme high dimensionality and collinearity of multi-omic datasets often make it challenging to interpret correlation networks. Hierarchical all-against-all association testing (HALLA) offers an alternative to traditional correlation network analyses by using a hierarchical hypothesis-testing framework and false discovery rate (FDR)

correction⁸⁵. HALLA clusters datasets hierarchically (thus handling collinearity) and then tests associations between clusters to identify both linear and nonlinear relationships across continuous and categorical data. In a study of the gut microbiome and metabolome in infants, HALLA validated previous findings regarding microbiome–metabolite associations and uncovered associations between *Prevotella* and inosine (a purine-derived metabolite), as well as between several bile acids and specific microbial genera⁸⁵.

Prediction of features in one omic based on another omic

Determining whether features in one omic dataset can successfully predict features in another could allow researchers to impute missing omic data. This objective is often addressed by various machine learning (ML) methods (Fig. 3, bottom right). As an example, MelonnPan, employs elastic net models to predict metabolite levels from taxonomic or functional microbial profiles⁸⁶. Applying MelonnPan to two datasets including over 200 individuals with Crohn's disease, ulcerative colitis (UC) and healthy controls, the authors demonstrated that their models resulted in significant correlations between predicted and observed faecal metabolic levels for over 50% of the identified metabolites⁸⁶. In another study, a meta-analysis of 10 human gut microbiome–metabolome datasets with a total of 1,733 samples from healthy participants, ML random forest models identified 97 metabolites that were 'robustly well predicted', that is, consistently well predicted across multiple datasets, spanning both well-characterized and novel microbial pathways⁸⁷.

As an alternative to evaluating how well features from one omic can be predicted by another omic, researchers can also estimate how much of the variance in these features can be explained by other omics. For example, a study investigating the efficacy of the IBD drug 5-aminosalicylic acid (5-ASA) used linear models with variance decomposition to show that the variance in 5-ASA-modulated metabolites was primarily explained by drug levels (-29%), followed by microbiome features (-19%), and other host factors such as diet and disease type⁸⁸.

Methods for linking omics to phenotypes

Methods for quantifying the relationships between various omic profiles and clinical phenotypes (for example, disease state or severity) aim to determine which profiles are most informative and how they can be integrated to improve disease prediction and biomarker discovery. They range from methods mapping broad associations and understanding overall predictive power (and ultimately diagnostic potential) to more refined analyses aimed at identifying specific multi-omic biomarkers and potential mechanistic interactions (Fig. 4).

Comparing different omics' associations with disease

Associations between each omic and a phenotype of interest can be assessed using methods previously described for measuring global associations between omics (Fig. 4, bottom left). For example, Mantel tests demonstrated a significant association between gut microbiome profiles and 11 clinical indices in a cohort of patients with Graves' disease⁸⁹. Other studies have used other statistical tests, such as permutational multivariate analysis of variance (PERMANOVA), to estimate which omic profile associates most strongly with a phenotype. The iHMP study used PERMANOVA to demonstrate that IBD status was significantly associated with faecal metabolomic profiles and colon biopsy RNA-seq profiles, but not with faecal microbiome profiles, probably due to the large variation between patients in disease severity and microbiome stability¹⁷.

Another approach for comparing global microbiome–phenotype associations is to train machine learning models to predict the phenotype using each omic separately and then compare model performance. In a study of non-alcoholic fatty liver disease (NAFLD), random forest models were trained to predict cirrhosis in a case-control cohort, using either gut microbiome species abundances or faecal metabolite levels as features⁹⁰. Interestingly, both models achieved a similar area under

the receiver operating characteristic curve (AUC), indicating that the disease signature of both omic profiles is similarly significant and that both omics could potentially be used for non-invasive diagnosis.

Improving disease prediction using multiple omics

Clearly, in some cases, a combination of omics can predict the phenotype of interest better than a single one. This can be assessed using integrative multi-omic ML models trained using several different integration approaches, often referred to as 'early', 'late' and 'intermediate' integration^{42,91}. In early integration, the most commonly used approach, different omics are first concatenated, forming a single ('wide') table of features per sample, and then ML pipelines are applied (hence this approach is also referred to as 'concatenation based'; Fig. 4, bottom middle). A model based on integrated microbiome (16S) and host (RNA-seq) omics, for example, significantly outperformed single-omic models in predicting future relapse in patients with IBD⁷. Another integrated microbiome–metabolome ML classifier (gradient boosting decision trees) of myalgic encephalomyelitis/chronic fatigue syndrome was shown to significantly outperform classification models based on a single profile across multiple ML algorithms⁴. Notably, in several other studies, multi-omic early integration models did not result in significantly improved disease prediction when compared to single omics, either due to high degrees of information shared between different omics, or due to one omic being particularly predictive of disease and masking the predictive ability of all others^{31,92–96}. Moreover, while early integration models are straightforward to implement, the extreme high dimensionality and heterogeneity of the concatenated table require careful feature selection and modelling choices. ML models trained on the concatenated omics are often biased towards omics with higher numbers of features, often necessitating balancing feature numbers via omic-specific feature selection.

Late integration, also known as 'stacked generalization', involves training ML models for each omic dataset independently, and then generating an ensemble model designed to provide final predictions⁹⁷. One study, for example, recommended combining microbiome risk scores with other omic-based risk scores by training a final logistic regression model on the different omic-specific risk predictions⁵. This combined risk-score approach led to significantly improved mortality prediction in a cohort of hospitalized patients with COVID-19. In another study, a stacked generalization model outperformed any single-omic model in predicting gestational age in a longitudinal cohort of 17 pregnant women⁶. Notably, recent methods such as cooperative learning attempt to bridge early and late integration by encouraging agreement between predictions derived from different omics within a single model⁹⁸.

Intermediate integration refers to various methods that first transform omics into a latent space, highlighting shared and distinct variation axes across omics (possibly incorporating disease labels in the process), and then use the latent variables to predict disease state. While intermediate integration methods can potentially be used for their classification ability, they are more often used for the identification of disease-related latent variables and are therefore discussed below.

Identifying multi-omic disease biomarkers

Moving from global to more specific associations, researchers may aim to identify specific subsets of features from all omics that shift or constitute a signature of disease. A straightforward step is to apply univariate statistical tests to each feature in each omic, identifying all features significantly associated with the phenotype (after correcting for multiple hypothesis testing). Univariate tests are commonly used, including tests that are specifically tailored to the characteristics of each omic; for example, DESeq2⁹⁹, MAST¹⁰⁰ or ALDEx2¹⁰¹ for transcriptomic data, or MaAsLin2¹⁰², metagenomeSeq¹⁰³ or ANCOM¹⁰⁴ for metagenomic data. Many of these tests allow to control for confounding variables, a crucial consideration in microbiome studies^{105,106}. While

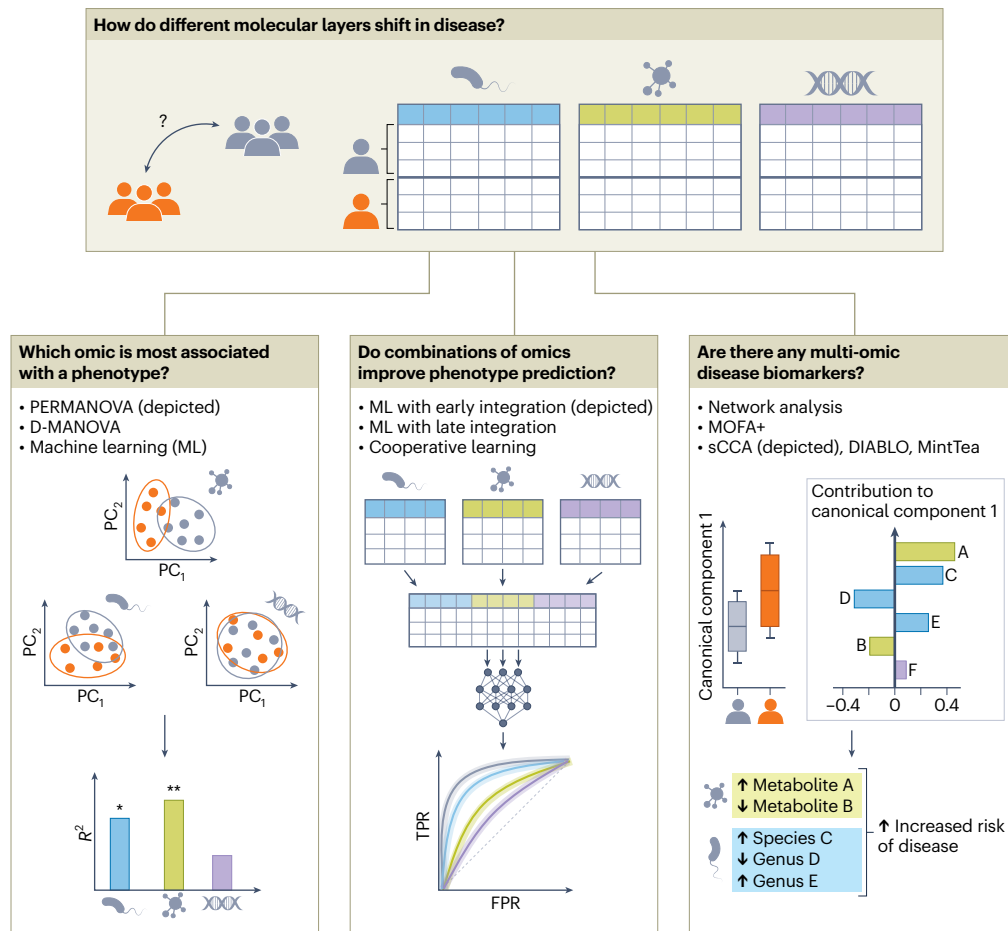


Fig. 4 | Multi-omic integration approaches to determine how different molecular layers shift in disease. Researchers could explore which omic profile presents the strongest association with a certain host disease state or phenotype (bottom left), assess whether combining omic profiles can result in improved predictive modelling of disease (bottom middle) or identify sets of features from

different omics that can collectively serve as a biomarker for disease (bottom right). Each box lists key methods that can be used to address the specific question, with the method illustrated in that box marked as 'depicted'. sCCA, sparse canonical correlation analysis.

this approach can provide a comprehensive list of potential markers, it does not consider the relationships between the different features and the different omics.

A possible approach to address this caveat is to combine the resulting list of disease-associated features, as identified above, with the previously described concept of correlation networks, as implemented in transkingdom network analysis (TkNA)¹⁰⁷. TkNA involves identifying disease-associated multi-omic features, calculating pairwise correlations between these features (within each study group), filtering out correlations that are not likely to reflect causal relationships (using 'correlation inequalities'), and then analysing the topology of the constructed networks to identify clusters or key regulatory nodes with a potential causal role in disease. Conveniently, TkNA supports a variety of study designs including a meta-analysis setup and has been previously applied to several studies of host–microbiome interactions in disease^{108–110}.

An alternative approach is to train multi-omic ML models using all omics concatenated together, but to further compute feature importance scores and identify the top contributing features serving as potential biomarkers. For example, a feature importance analysis identified a combination of different SNPs, including one in the pro-inflammatory gene *IL23R*, together with several *Bacteroides* species, as the most informative predictors of future relapse in patients with UC⁷. This interpretation, however, requires caution as (1) feature importance estimations are only meaningful for well-performing models; (2) feature selection methods, when employed, may remove

features that are redundant to the model, yet biologically important; and (3) different ML models and different feature importance metrics may behave differently in the presence of correlated (that is, redundant) features, either randomly assigning one feature as important and the other as non-important (introducing instability to feature scores), or reducing the importance scores of all correlated features. See recent publications^{111,112} for more detail.

The intermediate integration approach described above is another approach for finding multi-omic disease signatures^{42,113,114}. Such methods typically search for a latent representation of the samples, based on all omics, that best explains the variation in the data, and specifically, the variation associated with a phenotype of interest (Fig. 4, bottom right). The previously described MOFA+ could be used to identify molecular variation that differentiates between study groups and was applied to lower airway microbiome and host transcriptome profiles from a cohort of preschool children¹¹⁵. It identified a latent factor, mapping a trajectory from health through recurrent wheeze to established asthma, driven primarily by an increase in *Haemophilus* and *Neisseria* gene abundances alongside a transcriptomic signature of increased IL13 and eosinophilia.

Data integration analysis for biomarker discovery using latent components (DIABLO)¹¹⁶ is another multi-omic intermediate integration method. It expands sparse generalized canonical correlation analysis (sgCCA) to a classification framework by searching for correlated information among multiple omics that also correlates with the

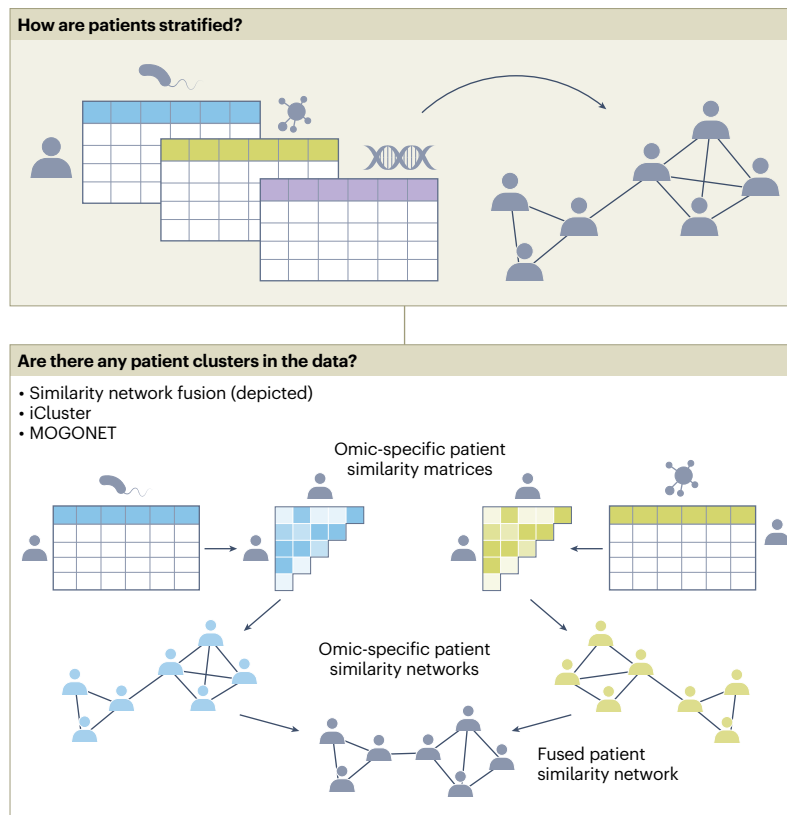


Fig. 5 | Multi-omic integration approaches for determining how patients are stratified on the basis of their different molecular profiles. When addressing this research question, researchers could specifically explore whether the data can be used to cluster patients into groups with similar molecular profiles.

The bottom box lists key methods that can be used to address this question, with the method illustrated marked as 'depicted'. MOGONET, Multi-Omics Graph Convolutional Networks.

phenotype of interest. DIABLO was used to find multi-omic signatures of brain damage in preterm infants using gut metagenomics, metabolomics, clinical, immunological and neurophysiological data¹¹⁷. Analysing the main DIABLO component, involving *Klebsiella* (increased), short-chain fatty acids (decreased) and levels of certain T cells, the authors hypothesized that *Klebsiella* species overgrowth may exacerbate brain injury by triggering changes in immunological development.

Inspired by DIABLO, we recently developed multi-omic integration tool for microbiome analysis (MintTea), which uses a repeated data subsampling approach to identify small multi-omic disease-associated modules that are also robust to data perturbations, noise and outliers¹¹⁸. These modules represent sets of features from different omics that are both significantly correlated with one another and collectively associated with the disease. Applied to gut microbiome and metabolome data from a cohort of patients with late-stage colorectal cancer, MintTea identified a module consisting of three bacterial species and four faecal metabolites that could classify diseased state.

When researchers hypothesize that the association between omic features and a phenotype is mediated by features of another omic¹¹⁹, mediation analysis can help decompose total effects (for example, associations) into direct and indirect (mediated) components. In microbiome studies, mediation analysis was used to estimate the indirect effect of diet on host health through modulation of certain gut bacteria^{120–122}, the effect of diet on stool or serum metabolites mediated by the gut microbiome^{16,19} or the indirect effect of compositional changes of the gut microbiome on host health via gut metabolites^{15,123,124}. A notable example is that of Yan et al.¹⁴, which used stepwise mediation analysis to implicate the airway microbiome in chronic obstructive pulmonary disease through effects on the airway metabolome and host gene expression, with key findings subsequently validated in a mouse model.

Methods for clustering multi-omic samples

Identification of subgroups (or 'clusters') of patients that share similar multi-omic profiles (Fig. 5) could provide insights into distinct disease manifestations with potentially different underlying mechanisms. In microbiome research, samples are occasionally grouped on the basis of their taxonomic profiles, leveraging established clustering techniques such as hierarchical clustering or *K*-means^{125,126}, or alternatively more innovative approaches tailored to the specific characteristics of microbiome data^{127–129}. Clustering multi-omic data, however, is more challenging as the different data characteristics of each omic must be accounted for. Researchers can use either adaptations of traditional clustering methods, or newer methods specifically designed for multi-omic datasets. Such algorithms can be classified into several primary strategies including similarity network fusion¹³⁰, latent variable models¹³¹ and spectral clustering¹³², with some additional methods specifically tailored for microbiome-related data^{133–136}.

Similarity network fusion (SNF) is a technique that creates a sample-similarity network based on each omic profile independently, and uses an iterative process to combine the different networks into one final 'consensus' network that can then be partitioned into clusters¹³⁰. As one example, Raita et al.¹³⁷ applied SNF to multi-omic data from infants with respiratory syncytial virus bronchiolitis, integrating clinical data as well as nasopharyngeal microbiome, transcriptome and metabolome data. Four distinct subgroups of infants with respiratory syncytial virus bronchiolitis were identified on the basis of clinical presentation, major bacterial species and immune response.

Spectral clustering transforms similarity graphs into simpler forms based on the eigenvalues and eigenvectors of the graph¹³⁸, facilitating the detection of natural clusters in complex and irregular data

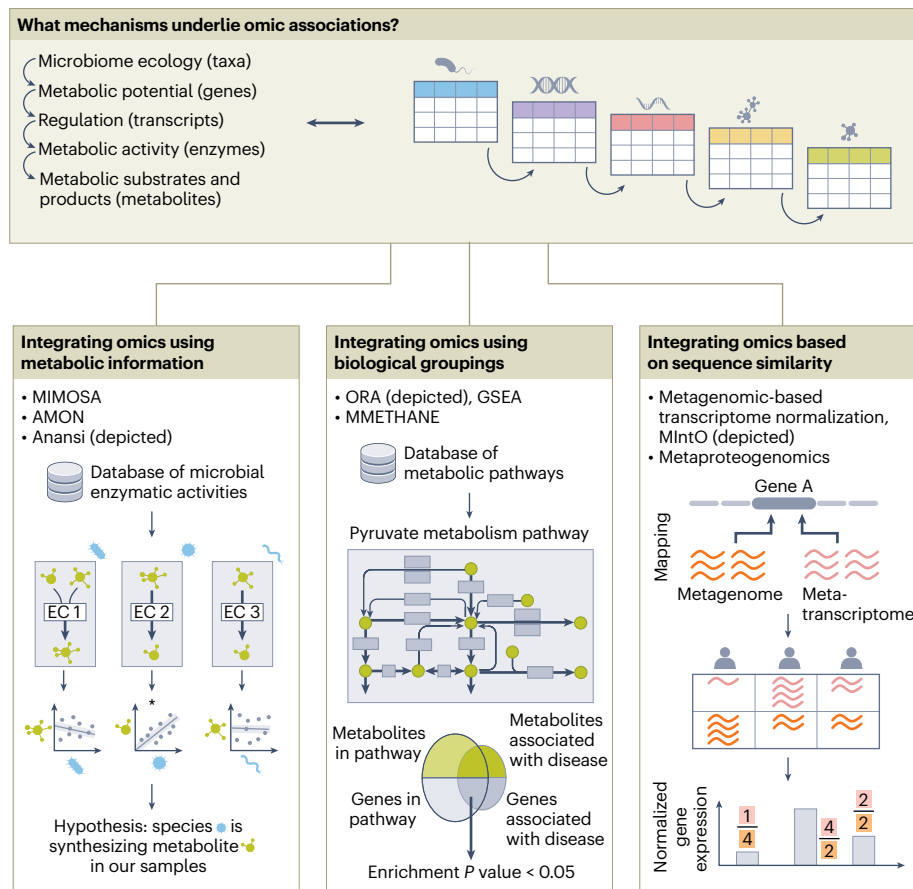


Fig. 6 | Multi-omic integration approaches for identifying putative mechanistic links between omics. Diverse databases can be used to construct microbial metabolic networks and to identify statistical associations supported by metabolic knowledge (bottom left), to group features based on biological information and identify significant associations between omics at different levels of abstraction (bottom middle), or to rely on sequence similarity, from DNA

to RNA to protein sequences, when interpreting links between metagenomes, metatranscriptomes and metaproteomes (bottom right). Each box lists key methods that can be used to address the specific question, with the method illustrated in that box marked as 'depicted'. EC, Enzyme Commission (enzyme identifier); ORA, over-representation analysis; GSEA, gene-set enrichment analysis; MIntO, microbiome integrated meta-omics.

and forming the basis for many multiview clustering algorithms^{139–141}. One such method is Spectrum¹³², which integrates multiple omics by emphasizing local sample similarities and reducing the impact of noisy or uneven data. It was used to analyse methylation, transcriptomic and microbiome data from head and neck squamous cell carcinoma tumours, identifying biologically relevant clusters that improved understanding of tumour heterogeneity¹⁴².

iCluster is another popular method based on a joint latent variable model that captures both shared and unique omic variation, followed by K -means for identifying clusters¹³¹. Unlike spectral clustering, which focuses on analysing similarity between samples (based on a single or multiple omics), iCluster, as well as its notable extensions iCluster+¹⁴³ and iClusterBayes¹⁴⁴, explicitly integrate the different omics into a shared space before clustering. iCluster+ was applied to The Cancer Genome Atlas Breast Invasive Carcinoma and clustered patients into groups with different survival conditions based on intratumour microbial abundance and metabolic pathway activity¹⁴⁵.

Recently, several methods have been developed specifically for microbiome data, including integrative stochastic variational variable selection (I-SVVS)¹³³, merged affinity network association clustering (MANAclust)¹³⁴, methods based on symmetric non-negative matrix factorization (SNMF)^{135,136} and non-negative matrix factorization with graph regularized optimal transport (NMFOT)¹⁴⁶. A comprehensive review focusing on clustering methods for multi-omic data more broadly is available¹⁴⁷.

Methods that utilize prior biological knowledge for multi-omic integration

Methods described so far are largely 'data-driven' and completely ignore large resources of existing biological knowledge such as KEGG^{148,149}, BioCyc^{150,151}, UniProt¹⁵², HMDB¹⁵³, eggNOG¹⁵⁴ and others^{155–157}, in which massive amounts of biological data have been carefully catalogued and curated by experts into relational databases. Each database typically includes millions of biological entities, many of which are annotated and mechanistically linked to other entities. This continuously growing treasure trove of biological knowledge¹⁵⁸ warrants the development of methods that can incorporate such knowledge into multi-omic analysis to generate easily interpretable and more biologically grounded hypotheses (Fig. 6).

Integrating omics using metabolic networks

Several knowledge-based multi-omic integration methods rely on well-characterized biochemical links between genes, enzymes, metabolic reactions and metabolites, which can be combined to construct large-scale networks of microbial metabolism^{149,151,155,156,159} (Fig. 6, bottom left). Methods integrating microbiome and metabolome data using these networks can be used to determine which metabolites are controlled by, or originate from, the microbial community. Model-based integration of metabolite observations and species abundances (MIMOSA), for example, computes the potential of a given microbial community to produce or consume each observed metabolite,

based on metagenomic data and metabolic modelling, and examines whether this inferred potential correlates with the observed levels of each metabolite^{25,160}. Annotation of metabolite origins via networks (AMON) receives a list of observed metabolites and uses genomic and metabolic reference information to predict whether the likely origin of each metabolite is the host or microbial enzymes¹⁶¹. An R package named 'Anansi' offers a convenient method for capturing correlations between features from different omics, while constraining the analysis only to pairs of features that are known to be mechanistically linked (based on biological databases)¹⁶².

Integrating omics using biologically meaningful groupings

Biologically meaningful groupings of molecular features, such as metabolic pathways defined in KEGG Pathways¹⁶³, MetaCyc¹⁶⁴ or WikiPathways¹⁶⁵, can be used to identify pathways perturbed in a certain disease across multiple omic layers (Fig. 6, bottom middle). For example, pathway-based integration of gene expression and metabolite data has been applied to study host–microbiome interactions in healthy and prediabetic participants¹⁶⁶. By integrating host transcripts, proteins, metabolites and cytokines, the authors first identified pathways altered during respiratory viral infection, and then the gut and nasal bacterial genera associated with each pathway. Importantly, in pathway-level analyses, different decisions about the reference database, the background feature set (that is, the 'universe' of features considered in the analysis), the *P*-value cut-offs and the weight given to features of different omics may all have a substantial effect on the pathways identified as important, necessitating careful sensitivity analysis^{167–169}.

Even within each individual omic profile, previous knowledge can be used to group features into biologically meaningful categories, helping to reduce data dimensionality and enhance both multi-omic integration and interpretability of results. Metabolites can be grouped into chemical classes, such as those listed in the Human Metabolome Database¹⁵⁹; genes and proteins can be grouped by Gene Ontology terms¹⁷⁰ or protein families as listed in UniProtKB¹⁵²; and microbial species abundances can be aggregated into higher taxonomic levels. In one proposed protocol for integrating metagenomic and metabolomic data, genes are grouped into KEGG functional modules as a previous step before testing module–metabolite associations¹⁷¹. Similarly, microbes and metabolites to host analysis engine (MMETHANE)¹⁷² uses a feedforward neural network to learn simple, easily interpretable rules for predicting disease state from paired microbiome–metabolome profiles, incorporating phylogenetic distances between microbial features and structure-based chemical similarities between metabolite features to ensure that the model learns only from biologically meaningful feature groupings.

Integrating omics based on sequence similarity

Sequence-centric integration methods represent another class of knowledge-based techniques¹⁷³ that are specifically relevant for the integration of metagenomics, metatranscriptomics and metaproteomics. One approach aligns metagenomic and metatranscriptomic reads to the same microbial gene catalogue, generating genomic and transcriptional profiles spanning the same set of genes (and taxa)^{8–10}. These paired profiles enable improved normalization of metatranscriptomic abundances by accounting for source species abundance or gene copy number (Fig. 6, bottom right). This helps distinguish transcripts that are rare due to low gene abundance from those that are genuinely lowly expressed¹⁰. Such analyses have revealed that several oral species that survive transit to the gut show limited transcriptional activity there²³, or were used to define a 'core' versus 'variable' metatranscriptome in healthy adults²⁴.

Integrating metagenomics and metaproteomics, often referred to as 'metaproteogenomics', can contribute to more comprehensive and reliable characterization of microbiome-associated proteomes^{11,12,174,175}.

In this approach, metagenomic data from the same samples are used to construct sample- or cohort-specific reference protein databases against which peptide spectra are matched. Such strategies have been implemented in several human microbiome studies^{11,176–179}. For example, this approach revealed that *Bacteroides vulgatus* produced excessive proteases in the gut of patients with UC, and that these microbial proteins were highly predictive of disease severity¹⁸⁰.

Conclusion

The rapid expansion of multi-omic integration methods can be overwhelming for microbiome researchers, making it challenging to determine which method is best suited for addressing a given research objective. This Review aims to address this challenge, guide microbiome researchers in selecting the most appropriate method for their specific scientific question and empower them to harness state-of-the-art integration tools more confidently.

Despite the promise of multi-omic integration methods, several challenges still hinder widespread application in microbiome research. First, the field lacks standardized benchmarking protocols for systematic method comparisons. Multi-omic benchmarking is complicated by data heterogeneity, the absence of ground-truth datasets and the difficulties in defining evaluation criteria. Second, the generalizability of multi-omic findings across datasets is often difficult to assess due to differences in sample processing, profiling technologies, population characteristics and study design, ultimately limiting reproducibility. These factors are known to impact single-omic analyses but are further amplified in multi-omic settings. Third, many integration methods produce complex outputs that are difficult to interpret. Incorporating prior knowledge, as discussed in this Review, can help contextualize findings and improve interpretability. Finally, many integration methods cannot easily be expanded to support more complex study designs, including specifically longitudinal datasets or studies with multiple confounding factors that need to be accounted for.

New methodological directions continue to emerge, aiming to address some of the above challenges. For example, frameworks such as timeOmics¹⁸¹, pipeline for the analysis of longitudinal multi-omics (PALM)¹⁸² and a method for the functional integration of spatial and temporal omics data (MEFISTO)¹⁸³ have been introduced to specifically integrate longitudinal multi-omic data and capture system-level temporal dynamics (see ref. 184). Artificial intelligence and deep learning approaches are also starting to play an important role in multi-omic integration owing to their ability to model high-dimensional and nonlinear relationships, and possibly extract meaningful patterns from raw and unstructured omic data (Supplementary Note 1).

As the field advances, increased attention should be given to the validation of integration results, their robustness to small perturbations, noise or parameter choices, and their reproducibility across different datasets, cohorts and study designs.

References

1. Gilbert, J. A. et al. Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
2. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* **14**, 508–522 (2016).
3. Marchesi, J. R. et al. The gut microbiota and host health: a new clinical frontier. *Gut* **65**, 330 (2016).
4. Xiong, R. et al. Multi-omics of gut microbiome-host interactions in short- and long-term myalgic encephalomyelitis/chronic fatigue syndrome patients. *Cell Host Microbe* **31**, 273–287.e5 (2023).
5. Wang, C. et al. Microbial risk score for capturing microbial characteristics, integrating multi-omics data, and predicting disease risk. *Microbiome* **10**, 121 (2022).

6. Ghaemi, M. S. et al. Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. *Bioinformatics* **35**, 95–103 (2019).
7. O'Sullivan, J. et al. Host–microbe multi-omics and succinotype profiling have prognostic value for future relapse in patients with inflammatory bowel disease. *Gut Microbes* **17**, 2450207 (2025).
8. Zhang, Y., Thompson, K. N., Huttenhower, C. & Franzosa, E. A. Statistical approaches for differential expression analysis in metatranscriptomics. *Bioinformatics* **37**, i34–i41 (2021).
9. Saenz, C., Nigro, E., Gunalan, V. & Arumugam, M. MIntO: a modular and scalable pipeline for microbiome metagenomic and metatranscriptomic data integration. *Front. Bioinform.* **2**, 846922 (2022).
10. Franzosa, E. A. et al. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* **13**, 360–372 (2015).
11. Valdés-Mas, R. et al. Metagenome-informed metaproteomics of the human gut microbiome, host, and dietary exposome uncovers signatures of health and inflammatory bowel disease. *Cell* **188**, 1062–1083.e36 (2025).
12. Singer, F., Kuhring, M., Renard, B. Y. & Muth, T. in *Proteogenomics: Methods and Protocols* (eds Allmer, J. & Kumar, A.) 297–318 (Springer, 2025); https://doi.org/10.1007/978-1-0716-4152-1_17
13. Priya, S. et al. Identification of shared and disease-specific host gene–microbiome associations across human diseases using multi-omic integration. *Nat. Microbiol.* **7**, 780–795 (2022).
14. Yan, Z. et al. Multi-omics analyses of airway host–microbe interactions in chronic obstructive pulmonary disease identify potential therapeutic interventions. *Nat. Microbiol.* **7**, 1361–1375 (2022).
15. Huang, Y. et al. Mapping the early life gut microbiome in neonates with critical congenital heart disease: multiomics insights and implications for host metabolic and immunological health. *Microbiome* **10**, 245 (2022).
16. Chen, L. et al. Influence of the microbiome, diet and genetics on inter-individual variation in the human plasma metabolome. *Nat. Med.* **28**, 2333–2343 (2022).
17. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
18. Price, N. D. et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat. Biotechnol.* **35**, 747–756 (2017).
19. Tang, Z.-Z. et al. Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. *Front. Genet.* **10**, 454 (2019).
20. Liu, C., Egor, S. & Hosein, M. A Metabolome- and metagenome-wide association network reveals microbial natural products and microbial biotransformation products from the human microbiota. *mSystems* <https://doi.org/10.1128/mSystems.00387-19> (2019).
21. Liang, L. et al. Distinct microbes, metabolites, and the host genome define the multi-omics profiles in right-sided and left-sided colon cancer. *Microbiome* **12**, 274 (2024).
22. Gosalbes, M. J. et al. Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE* **6**, e17447 (2011).
23. Franzosa, E. A. et al. Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl Acad. Sci. USA* **111**, E2329–E2338 (2014).
24. Abu-Ali, G. S. et al. Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nat. Microbiol.* **3**, 356–366 (2018).
25. Noecker, C. et al. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* **1**, e00013-15 (2016).
26. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
27. Asnicar, F. et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).
28. Bergström, G. et al. The Swedish CARDioPulmonary BioImage Study: objectives and design. *J. Intern. Med.* **278**, 645–659 (2015).
29. ter Horst, R. et al. Host and environmental factors influencing individual human cytokine responses. *Cell* **167**, 1111–1124.e13 (2016).
30. Verdi, S. et al. TwinsUK: the UK adult twin registry update. *Twin Res. Hum. Genet.* **22**, 523–529 (2019).
31. Fromentin, S. et al. Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nat. Med.* **28**, 303–314 (2022).
32. Nyholm, L. et al. Holo-Omics: integrated host–microbiota multi-omics for basic and applied biological research. *iScience* **23**, 101414 (2020).
33. Hernández-Lemus, E. & Ochoa, S. Methods for multi-omic data integration in cancer research. *Front. Genet.* **15**, 1425456 (2024).
34. Mohr, A. E., Ortega-Santos, C. P., Whisner, C. M., Klein-Seetharaman, J. & Jasbi, P. Navigating challenges and opportunities in multi-omics integration for personalized healthcare. *Biomedicine* **12**, 1496 (2024).
35. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* **14**, 1177932219899051 (2020).
36. Gomez-Cabrero, D. et al. Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* **8**, 11 (2014).
37. Flores, J. E. et al. Missing data in multi-omics integration: recent advances through artificial intelligence. *Front. Artif. Intell.* **6**, 1098308 (2023).
38. Yu, Y. et al. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biol.* **24**, 201 (2023).
39. Tarazona, S., Arzalluz-Luque, A. & Conesa, A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat. Comput. Sci.* **1**, 395–402 (2021).
40. Misra, B. B., Langefeld, C., Olivier, M. & Cox, L. A. Integrated omics: tools, advances and future approaches. *J. Mol. Endocrinol.* **62**, R21–R45 (2019).
41. Cantini, L. et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
42. Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).
43. Lamichhane, S., Sen, P., Dickens, A. M., Orešič, M. & Bertram, H. C. Gut metabolome meets microbiome: a methodological perspective to understand the relationship between host and microbiome. *Methods* **149**, 3–12 (2018).
44. Bauermeister, A., Mannocho-Russo, H., Costa-Lotufo, L. V., Jarmusch, A. K. & Dorrestein, P. C. Mass spectrometry-based metabolomics in microbiome investigations. *Nat. Rev. Microbiol.* **20**, 143–160 (2022).
45. Chong, J. & Xia, J. Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites* **7**, 62 (2017).
46. Doran, S. et al. Multi-omics approaches for revealing the complexity of cardiovascular disease. *Brief. Bioinform.* **22**, bbab061 (2021).

47. Xu, J. & Yang, Y. Gut microbiome and its meta-omics perspectives: profound implications for cardiovascular diseases. *Gut Microbes* **13**, 1936379 (2021).
48. Valles-Colomer, M. et al. Cardiometabolic health, diet and the gut microbiome: a meta-omics perspective. *Nat. Med.* **29**, 551–561 (2023).
49. Zhang, N., Kandalai, S., Zhou, X., Hossain, F. & Zheng, Q. Applying multi-omics toward tumor microbiome research. *iMeta* **2**, e73 (2023).
50. Huang, H., Vangay, P., McKinlay, C. E. & Knights, D. Multi-omics analysis of inflammatory bowel disease. *Immunol. Lett.* **162**, 62–68 (2014).
51. Jiang, D. et al. Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. *Front. Genet.* **10**, 995 (2019).
52. Chetty, A. & Blekhan, R. Multi-omic approaches for host-microbiome data integration. *Gut Microbes* **16**, 2297860 (2024).
53. Sankaran, K. & Holmes, S. P. Multitable methods for microbiome data integration. *Front. Genet.* **10**, 627 (2019).
54. Liu, Z. et al. Network analyses in microbiome based on high-throughput multi-omics data. *Brief. Bioinform.* **22**, 1639–1655 (2021).
55. Athieniti, E. & Spyrou, G. M. A guide to multi-omics data collection and integration for translational medicine. *Comput. Struct. Biotechnol. J.* **21**, 134–149 (2023).
56. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *ISME J.* **5**, 169–172 (2011).
57. Stanimirova, I. & Daszykowski, M. in *Comprehensive Analytical Chemistry* Vol. 82 (eds Jaumot, J. et al.) 227–264 (Elsevier, 2018).
58. Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D. & Lu, L. J. Computational and statistical analysis of metabolomics data. *Metabolomics* **11**, 1492–1513 (2015).
59. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
60. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
61. Smouse, P. E., Long, J. C. & Sokal, R. R. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Biol.* **35**, 627–632 (1986).
62. Székely, G. J. & Rizzo, M. L. The distance correlation t-test of independence in high dimension. *J. Multivar. Anal.* **117**, 193–213 (2013).
63. Gower, J. C. Generalized procrustes analysis. *Psychometrika* **40**, 33–51 (1975).
64. Amir, T., Kovalsky, S. & Dym, N. Symmetrized robust procrustes: constant-factor approximation and exact recovery. Preprint at <https://arxiv.org/abs/2207.08592> (2022).
65. García-Pérez, A. & Cabrero-Ortega, M. A. Robust morphometric analysis based on landmarks. Applications. Preprint at <https://arxiv.org/abs/1703.04642> (2017).
66. Deek, R. A., Ma, S., Lewis, J. & Li, H. Statistical and computational methods for integrating microbiome, host genomics, and metabolomics data. *eLife* **13**, e88956 (2024).
67. Wold, H. in *Multivariate Analysis* (ed. Krishnajah, P. R.) 391–420 (Academic Press, 1966).
68. Hotelling, H. Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936).
69. Dolédec, S. & Chessel, D. Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshw. Biol.* **31**, 277–294 (1994).
70. Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7**, 523–542 (2013).
71. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
72. Ingham, A. C. et al. Specific gut microbiome members are associated with distinct immune markers in pediatric allogeneic hematopoietic stem cell transplantation. *Microbiome* **7**, 131 (2019).
73. Schwartz, S. et al. A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biol.* **13**, r32 (2012).
74. Heintz-Buschart, A. et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2**, 16180 (2016).
75. Kostic, A. D. et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).
76. Helmer, M. et al. On the stability of canonical correlation analysis and partial least squares with application to brain–behavior associations. *Commun. Biol.* **7**, 217 (2024).
77. McIntosh, A. R. Comparison of canonical correlation and partial least squares analyses of simulated and empirical data. Preprint at <https://arxiv.org/abs/2107.06867> (2021).
78. Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735 (2021).
79. Marabita, F. et al. Multiomics and digital monitoring during lifestyle changes reveal independent dimensions of human biology and health. *Cell Syst.* **13**, 241–255.e7 (2022).
80. Wilmsanski, T. et al. Blood metabolome predicts gut microbiome α -diversity in humans. *Nat. Biotechnol.* **37**, 1217–1228 (2019).
81. Fackelmann, G. et al. Gut microbiome signatures of vegan, vegetarian and omnivore diets and associated health outcomes across 21,561 individuals. *Nat. Microbiol.* **10**, 41–52 (2025).
82. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17 (2005).
83. Albrich, W. C. et al. A high-risk gut microbiota configuration associates with fatal hyperinflammatory immune and metabolic responses to SARS-CoV-2. *Gut Microbes* **14**, 2073131 (2022).
84. Guan, S.-W., Lin, Q., Wu, X.-D. & Yu, H.-B. Weighted gene coexpression network analysis and machine learning reveal oncogene associated microbiome plays an important role in tumor immunity and prognosis in pan-cancer. *J. Transl. Med.* **21**, 537 (2023).
85. Ghazi, A. R. et al. High-sensitivity pattern discovery in large, paired multiomic datasets. *Bioinformatics* **38**, i378–i385 (2022).
86. Mallick, H. et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* **10**, 3136 (2019).
87. Muller, E., Algavi, Y. M. & Borenstein, E. A meta-analysis study of the robustness and universality of gut microbiome–metabolome associations. *Microbiome* **9**, 203 (2021).
88. Mehta, R. S. et al. Gut microbial metabolism of 5-ASA diminishes its clinical efficacy in inflammatory bowel disease. *Nat. Med.* **29**, 700–709 (2023).
89. Zhu, Q. et al. Compositional and genetic alterations in Graves’ disease gut microbiome reveal specific diagnostic biomarkers. *ISME J.* **15**, 3399–3411 (2021).
90. Oh, T. G. et al. A universal gut-microbiome-derived signature predicts cirrhosis. *Cell Metab.* **32**, 878–888.e6 (2020).
91. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv.* **49**, 107739 (2021).
92. Cantoni, C. et al. Alterations of host–gut microbiome interactions in multiple sclerosis. *eBioMedicine* **76**, 103798 (2022).

93. Franzosa, E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
94. Bokulich, N. A. et al. Multi-omics data integration reveals metabolome as the top predictor of the cervicovaginal microenvironment. *PLoS Comput. Biol.* **18**, e1009876 (2022).
95. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
96. Ning, L. et al. Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts. *Nat. Commun.* **14**, 7135 (2023).
97. Naimi, A. I. & Balzer, L. B. Stacked generalization: an introduction to super learning. *Eur. J. Epidemiol.* **33**, 459–464 (2018).
98. Ding, D. Y., Li, S., Narasimhan, B. & Tibshirani, R. Cooperative learning for multiview analysis. *Proc. Natl Acad. Sci. USA* **119**, e2202113119 (2022).
99. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
100. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
101. Fernandes, A. D. et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).
102. Mallick, H. et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput. Biol.* **17**, e1009442 (2021).
103. Paulson, J. N., Colin Stine, O., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
104. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
105. Vujkovic-Cvijin, I. et al. Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
106. Forslund, S. K. et al. Combinatorial, additive and dose-dependent drug–microbiome associations. *Nature* **600**, 500–505 (2021).
107. Newman, N. K. et al. Transkingdom Network Analysis (TkNA): a systems framework for inferring causal factors underlying host–microbiota and other multi-omic interactions. *Nat. Protoc.* **19**, 1750–1778 (2024).
108. Rodrigues, R. R. et al. Transkingdom interactions between Lactobacilli and hepatic mitochondria attenuate western diet-induced diabetes. *Nat. Commun.* **12**, 101 (2021).
109. McCulloch, J. A. et al. Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1. *Nat. Med.* **28**, 545–556 (2022).
110. Morgun, A. et al. Uncovering effects of antibiotics on the host and microbiota using transkingdom gene networks. *Gut* **64**, 1732–1743 (2015).
111. Chen, V. et al. Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. *Nat. Methods* **21**, 1454–1461 (2024).
112. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: a unified framework for machine learning interpretability. Preprint at <https://arxiv.org/abs/1909.09223> (2019).
113. Wang, Y. et al. Multi-modal intermediate integrative methods in neuropsychiatric disorders: a review. *Comput. Struct. Biotechnol. J.* **20**, 6149–6162 (2022).
114. Xie, Z. et al. Integrated multi-omics analysis reveals gut microbiota dysbiosis and systemic disturbance in major depressive disorder. *Psychiatry Res.* **334**, 115804 (2024).
115. Macowan, M. et al. Deep multiomic profiling reveals molecular signatures that underpin preschool wheeze and asthma. *J. Allergy Clin. Immunol.* **155**, 94–106 (2025).
116. Singh, A. et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).
117. Seki, D. et al. Aberrant gut–microbiota–immune–brain axis development in premature neonates with brain damage. *Cell Host Microbe* **29**, 1558–1572.e6 (2021).
118. Muller, E., Shiryay, I. & Borenstein, E. Multi-omic integration of microbiome data for identifying disease-associated modules. *Nat. Commun.* **15**, 2621 (2024).
119. Jiang, H. et al. Multimedia: multimodal mediation analysis of microbiome data. *Microbiol. Spectr.* **13**, e01131-24 (2024).
120. Shi, H. et al. The gut microbiome as mediator between diet and its impact on immune function. *Sci. Rep.* **12**, 5149 (2022).
121. Serrano, D. et al. Microbiome as mediator of diet on colorectal cancer risk: the role of vitamin D, markers of inflammation and adipokines. *Nutrients* **13**, 363 (2021).
122. Zhu, S. et al. The gut microbiome in subclinical atherosclerosis: a population-based multiphenotype analysis. *Rheumatology* **61**, 258–269 (2021).
123. Chen, J. et al. Plasma metabolites as mediators between gut microbiota and Parkinson’s disease: insights from Mendelian randomization. *Mol. Neurobiol.* <https://doi.org/10.1007/S12035-025-04765-0/TABLES/1> (2025).
124. Menni, C. et al. Serum metabolites reflecting gut microbiome alpha diversity predict type 2 diabetes. *Gut Microbes* **11**, 1632–1642 (2020).
125. Shi, Y., Zhang, L., Peterson, C. B., Do, K.-A. & Jenq, R. R. Performance determinants of unsupervised clustering methods for microbiome data. *Microbiome* **10**, 25 (2022).
126. Tan-Torres, A. L., Brooks, J. P., Singh, B. & Seashols-Williams, S. Machine learning clustering and classification of human microbiome source body sites. *Forensic Sci. Int.* **328**, 111008 (2021).
127. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7**, e30126 (2012).
128. Subedi, S., Neish, D., Bak, S. & Feng, Z. Cluster analysis of microbiome data by using mixtures of Dirichlet–multinomial regression models. *J. R. Stat. Soc. C* **69**, 1163–1187 (2020).
129. Fang, Y. & Subedi, S. Clustering microbiome data using mixtures of logistic normal multinomial models. *Sci. Rep.* **13**, 14758 (2023).
130. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
131. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
132. John, C. R., Watson, D., Barnes, M. R., Pitzalis, C. & Lewis, M. J. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics* **36**, 1159–1166 (2020).
133. Dang, T. et al. I-SVVS: integrative stochastic variational variable selection to explore joint patterns of multi-omics microbiome data. *Brief. Bioinform.* **26**, bbaf132 (2025).
134. Tyler, S. R. et al. Merged affinity network association clustering: joint multi-omic/clinical clustering to identify disease endotypes. *Cell Rep.* **35**, 108975 (2021).
135. Jiang, X., Hu, X. & Xu, W. Microbiome data representation by joint nonnegative matrix factorization with Laplacian regularization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **14**, 353–359 (2017).
136. Ma, Y., Zhao, J. & Ma, Y. MHSNMF: multi-view Hessian regularization based symmetric nonnegative matrix factorization for microbiome data analysis. *BMC Bioinform.* **21**, 234 (2020).

137. Raita, Y. et al. Integrated omics endotyping of infants with respiratory syncytial virus bronchiolitis and risk of childhood asthma. *Nat. Commun.* **12**, 3601 (2021).
138. Ng, A., Jordan, M. & Weiss, Y. On spectral clustering: analysis and an algorithm. In *NIPS'01: Proc. 15th International Conference on Neural Information Processing Systems: Natural and Synthetic* (eds Dietterich, T. et al.) 849–856 (MIT Press, 2001).
139. Xia, R., Pan, Y., Du, L. & Yin, J. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 28 <https://doi.org/10.1609/aaai.v28i1.8950> (AAAI Press, 2014).
140. de Sa, V. R. Spectral clustering with two views. In *ICML Workshop on Learning With Multiple Views* 20–27 (ICML, 2005).
141. Zhou, D. & Burges, C. J. C. Spectral clustering and transductive learning with multiple views. In *Proc. 24th International Conference on Machine Learning* 1159–1166 (Association for Computing Machinery, 2007).
142. Bard, J. E., Nowak, N. J., Buck, M. J. & Sinha, S. Multimodal dimension reduction and subtype classification of head and neck squamous cell tumors. *Front. Oncol.* **12**, 892207 (2022).
143. Mo, Q. et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl Acad. Sci. USA* **110**, 4245–4250 (2013).
144. Mo, Q. et al. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19**, 71–86 (2018).
145. Chen, F. et al. Integrating bulk and single-cell RNA sequencing data reveals the relationship between intratumor microbiome signature and host metabolic heterogeneity in breast cancer. *Front. Immunol.* **14**, 1140995 (2023).
146. Ma, Y. & Liu, L. NMFOT: a multi-view learning framework for the microbiome and metabolome integrative analysis with optimal transport plan. *npj Biofilms Microbiomes* **10**, 135 (2024).
147. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* **46**, 10546–10562 (2018).
148. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
149. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes And Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
150. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **42**, D459–D471 (2014).
151. Karp, P. D. et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **20**, 1085–1093 (2019).
152. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
153. David, S., Wishart et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
154. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
155. Heinken, A. et al. Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nat. Biotechnol.* **41**, 1320–1331 (2023).
156. Cheng, L. et al. gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Res.* **50**, D795–D800 (2022).
157. Wishart, D. S. et al. MiMeDB: the human microbial metabolome database. *Nucleic Acids Res.* **51**, D611–D620 (2023).
158. Rigden, D. J. & Fernández, X. M. The 2025 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.* **53**, D1–D9 (2025).
159. Wishart, D. S. et al. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2022).
160. Noecker, C., Eng, A., Muller, E. & Borenstein, E. MIMOSA2: a metabolic network-based tool for inferring mechanism-supported relationships in microbiome–metabolome data. *Bioinformatics* **38**, 1615–1623 (2022).
161. Shaffer, M. et al. AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data. *BMC Bioinform.* **20**, 614 (2019).
162. Bastiaanssen, T. F. S., Quinn, T. P. & Cryan, J. F. Knowledge-based integration of multi-omic datasets with Anansi: annotation-based analysis of specific interactions. Preprint at <https://arxiv.org/abs/2305.10832v1> (2023).
163. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
164. Caspi, R. et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **34**, D511–D516 (2006).
165. Martens, M. et al. WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).
166. Zhou, W. et al. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **569**, 663–671 (2019).
167. Wieder, C. et al. Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. *PLoS Comput. Biol.* **17**, e1009105 (2021).
168. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375 (2012).
169. Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J. J. Transcriptomic and metabolomic data integration. *Brief. Bioinform.* **17**, 891–901 (2016).
170. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
171. Pedersen, H. K. et al. A computational framework to integrate high-throughput ‘-omics’ datasets for the identification of potential mechanistic links. *Nat. Protoc.* **13**, 2781–2800 (2018).
172. Dawkins, J. J. & Gerber, G. K. MMETHANE: interpretable AI for predicting host status from microbial composition and metabolomics data. *Microbiome* **14**, 21 (2025).
173. Heintz-Buschart, A. & Westerhuis, J. A. A beginner’s guide to integrating multi-omics data from microbial communities. *Biochemistry* **44**, 23–29 (2022).
174. Schiebenhoefer, H. et al. Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteomic data analysis. *Expert Rev. Proteomics* **16**, 375–390 (2019).
175. Kumar, D., Yadav, A. K. & Dash, D. in *Proteome Bioinformatics* (eds Keerthikumar, S. & Mathivanan, S.) 17–29 (Springer, 2017).
176. Zhang, Y. et al. Discovery of bioactive microbial gene products in inflammatory bowel disease. *Nature* **606**, 754–760 (2022).
177. Erickson, A. R. et al. Integrated metagenomics/metaproteomics reveals human host–microbiota signatures of Crohn’s disease. *PLoS ONE* **7**, e49138 (2012).
178. Wang, T., Li, L., Figeys, D. & Liu, Y.-Y. Pairing metagenomics and metaproteomics to characterize ecological niches and metabolic essentiality of gut microbiomes. *ISME Commun.* **4**, ycae063 (2024).
179. Mills, R. H. et al. Evaluating metagenomic prediction of the metaproteome in a 4.5-year study of a patient with Crohn’s disease. *mSystems* <https://doi.org/10.1128/msystems.00337-18> (2019).
180. Mills, R. H. et al. Multi-omics analyses of the ulcerative colitis gut microbiome link *Bacteroides vulgatus* proteases with disease severity. *Nat. Microbiol.* **7**, 262–276 (2022).

181. Bodein, A., Scott-Boyer, M.-P., Perin, O., Lê Cao, K.-A. & Droit, A. timeOmics: an R package for longitudinal multi-omics data integration. *Bioinformatics* **38**, 577–579 (2022).
182. Ruiz-Perez, D. et al. Dynamic Bayesian networks for integrating multi-omics time series microbiome data. *mSystems* <https://doi.org/10.1128/msystems.01105-20> (2021).
183. Velten, B. et al. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* **19**, 179–186 (2022).
184. Sherwani, M. K. et al. Multi-omics time-series analysis in microbiome research: a systematic review. *Brief. Bioinform.* **26**, bbaf502 (2025).

Acknowledgements

We thank the Borenstein lab members for helpful feedback and insightful discussions during the process of writing this review. We also thank the many authors whose work inspired this review article, including those we could not cite due to space constraints. This work was supported in part by National Institutes of Health grant U19AG057377, Israel Science Foundation grant 2266/25 (to E.B.), and by Len Blavatnik and the Blavatnik Family Foundation. E.M. and T.B. were supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

Author contributions

E.M., T.B. and E.B. conceptualized the review and wrote the paper. E.M. and T.B. mapped existing multi-omic integration methods and synthesized the findings.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41564-026-02328-0>.

Correspondence and requests for materials should be addressed to Elhanan Borenstein.

Peer review information *Nature Microbiology* thanks Nathan Price and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2026