



## Multi-omic approaches for host-microbiome data integration

Ashwin Chetty <sup>a</sup> and Ran Blekhman <sup>b</sup>

<sup>a</sup>Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, IL, USA; <sup>b</sup>Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA

### ABSTRACT

The gut microbiome interacts with the host through complex networks that affect physiology and health outcomes. It is becoming clear that these interactions can be measured across many different omics layers, including the genome, transcriptome, epigenome, metabolome, and proteome, among others. Multi-omic studies of the microbiome can provide insight into the mechanisms underlying host-microbe interactions. As more omics layers are considered, increasingly sophisticated statistical methods are required to integrate them. In this review, we provide an overview of approaches currently used to characterize multi-omic interactions between host and microbiome data. While a large number of studies have generated a deeper understanding of host-microbiome interactions, there is still a need for standardization across approaches. Furthermore, microbiome studies would also benefit from the collection and curation of large, publicly available multi-omics datasets.

### ARTICLE HISTORY

Received 19 July 2023  
Revised 13 December 2023  
Accepted 18 December 2023

### KEYWORDS

Multimics; microbiome; network; disease; analysis; inference; host-microbiome interactions

### Introduction

The human gut microbiome comprises myriad bacteria, viruses, fungi, and protozoa that interact with each other as well as their host, ultimately affecting host physiology and disease. The microbiota contribute to various functions important for health, including digestion, immune regulation, maintenance of the intestinal mucosal barrier, and protection from pathogens.<sup>1</sup> Alterations in the microbiota have been associated with health conditions including colorectal cancer,<sup>2</sup> inflammatory bowel disease,<sup>3</sup> obesity,<sup>4</sup> and depression.<sup>5</sup>



Despite its role in health, it is only with recent advances in sequencing technology that researchers have been able to gain a holistic view of the composition of the gut microbiome. Sequencing of the gut metagenome has allowed for the discovery of species that were previously difficult to culture *in vitro*.<sup>6</sup> Although progress has been made in culturing bacteria previously thought to be “unculturable”,<sup>7</sup> a large number of taxa still have not been cultured,<sup>8</sup> and next-generation sequencing remains the standard for microbiome profiling.

Analysis of the microbiome has revealed a complex web of interactions between

microorganisms and their host. For example, microbes interact with one another in networks that vary spatially and with disease states.<sup>9</sup> Furthermore, analyses using metabolomics and metagenomics have demonstrated the ability of microbes to metabolize drugs,<sup>10</sup> and joint analyses of the microbiome and host transcriptome have also shown that the microbiota can regulate host gene expression.<sup>11</sup> Additional studies have identified associations between host gene expression and gut microbiome composition across disease states, underscoring the relevance of multi-omics analyses of the microbiome.<sup>12–14</sup>

### Current approaches to microbiome data analysis

Current approaches to microbiome community profiling usually involve either shotgun metagenomic or 16S rRNA amplicon sequencing.<sup>15</sup> In shotgun sequencing, DNA is first extracted from all of the cells in a sample, cleaved into small fragments, and sequenced. Computational methods are then used to align the reads against reference genomes or marker genes to infer the abundances of

**CONTACT** Ran Blekhman  [blekhman@uchicago.edu](mailto:blekhman@uchicago.edu)  Section of Genetic Medicine, Department of Medicine, The University of Chicago, 900 E. 57th St., Chicago, IL 60637, USA

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

taxa present in the sample.<sup>15</sup> Shotgun sequencing may also be performed with higher read depths in order to assemble bacterial genomes *de novo*. In 16S rRNA amplicon sequencing, only a segment of the 16S rRNA gene of bacterial genomes in a sample is amplified and sequenced. The 16S rRNA gene contains both conserved and variable regions. The conserved regions are used as a target for PCR primers, while the variable regions are used to determine the identities and abundances of taxa present in the sample.

While microbiome research has largely focused on bacteria, which comprise more than 99% of the genetic material in the microbiome, viral, fungal, and archaeal microbial components are also of interest.<sup>16</sup> The virome, for example, includes viruses that infect host cells and bacteriophages that target bacteria. Alterations in the virome have been associated with human diseases such as SARS-CoV-2, inflammatory bowel disease, diabetes, and colorectal cancer.<sup>17</sup> Because viruses lack universally conserved sequences such as the 16S region, viral sequencing is usually performed using shotgun or long-read sequencing. Viral genetic material is usually present in lower concentrations in samples, which may motivate the use of higher read depths when sequencing.<sup>17</sup> The fungal microbiota, as with viruses, are affected by host factors such as diet and antibiotic use.<sup>16</sup> Fungi have been shown to exhibit complex interactions with the host and with other members of the microbiome.<sup>18</sup> For example, there is evidence that fungi and bacteria together contribute to early-life immune education, and *Saccharomyces boulardii* has been used as a probiotic to treat bacterial infections.<sup>18</sup> Fungi also play a role in host health and immunity. *Candida albicans*, for example, has been shown to modulate Th17 helper T cells.<sup>16</sup> *Candida* species have also been associated with health conditions such as inflammatory bowel disease and liver disease.<sup>16</sup> Similar to bacteria, fungal sequencing can also be performed either with shotgun sequencing or through amplicon sequencing of the 18S, ITS1 and ITS2 regions.<sup>16</sup> However, reference ITS sequences for many fungal genomes are currently lacking, limiting the ability of sequencing projects to study the fungal microbiota, and motivating the development of alternate

classification strategies.<sup>19</sup> In contrast with other components of the microbiome, archaea represent members of the microbial community that have thus far received little attention, partially due to the lack of common archaeal pathogens. While there is evidence that archaea interact with other microbial components and form stable enterotypes, there is currently little evidence of associations between archaea and human health.<sup>20</sup> Like bacteria, archaea are quantified with either shotgun or 16S amplicon sequencing.<sup>21</sup>

After sequencing is performed and data is processed, microbial abundances can be represented as a two-dimensional matrix of counts where each cell represents the estimated abundance of a taxon present in a particular sample. Computational methods are then used to gain biological insights from the data. A common analysis pipeline involves identifying differentially abundant taxa between different treatment groups (such as in microbiome case-control studies), which may be performed using software such as EdgeR.<sup>22</sup> Many bioinformatics software packages can be downloaded with Bioconductor<sup>23</sup> in R or with Anaconda in Python, and Knight et al. provide a comprehensive overview of current best practices in microbiome analysis.<sup>24</sup>

## Omics layers analyzed with microbiome data

A multitude of different data layers can be analyzed along with the microbiome, such as genomics, transcriptomics, epigenomics, radiomics, proteomics, metabolomics, diet, and clinical outcomes. Each layer must be approached with its own considerations, which we review in this section.

### Host transcriptomics

Host transcriptomics can provide insights into the functional interactions between host genes and microbiome by allowing researchers to quantify gene expression activity across different treatments or disease states. Transcriptomics may be performed with short- or long-read sequencing.<sup>25</sup> Transcriptomics protocols usually filter out highly-abundant ribosomal rRNA prior to sequencing by either depleting rRNA or performing poly(A) selection.<sup>25</sup> After sequencing is performed, reads are filtered based on

their sequencing quality, aligned to reference genomes, and normalized for biases due to transcript length and compositionality.<sup>25</sup>

### **Metatranscriptomics**

Metatranscriptomic analysis techniques allow researchers to quantify the abundances of microbial gene transcripts in a sample, which can provide insight into the functional characteristics of the microbiota.<sup>26</sup> Metatranscriptomics experimental protocols differ based on the organisms being studied. For example, prokaryotic mRNA lacks a poly-adenine tail, which precludes the use of poly(A) selection.<sup>26</sup> After next-generation sequencing is performed, transcripts are aligned to metatranscriptomic reference genomes and quantified.<sup>26</sup>

### **Host genetics**

Studies combining host genetics with microbiome measurements have demonstrated that microbiome phenotypes are heritable.<sup>27</sup> As a result, studies may be interested in unraveling the host genetic determinants of microbiome composition. A variety of technologies exist for performing genotyping, including next-generation whole genome sequencing, whole exome sequencing, and genotyping arrays. Kockum et al. provide an in-depth overview of current genotyping strategies,<sup>28</sup> and a comprehensive discussion on the influence of host genetics and the gut microbiome is offered by Goodrich et al.<sup>29</sup>

### **Metabolomics**

Metabolomics allows researchers a better understanding of the biochemical and metabolic processes that involve the host and microbiota. Metabolomics data is usually generated using mass spectrometry, where molecules are identified by cross-referencing their mass-to-charge ratios against reference databases. The direct measurement of small molecules provides information about cellular processes and can point to

mechanistic interactions between host and microbiota. Chong et al. provide a review of current techniques used to perform integrative analyses with metabolomic and microbiome data.<sup>30</sup>

### **Metaproteomics**

Metaproteomic analyses quantify the proteins produced by host and microbiome, which can provide insight into the functional role of microorganisms in host health. Complementary to metagenomics and metatranscriptomics, metaproteomic measurements reflect the activity of cellular translational and post-translational processes.<sup>31</sup> As with metabolomics, metaproteomics is generally performed using mass spectrometry. Proteomics database lookup is computationally expensive, and researchers should be mindful that the results of proteomic analyses may be sensitive to the particular choice of mass spectra database used.<sup>31</sup> Peters et al. provide an in-depth review on the role of proteomics analysis in microbiome studies.<sup>31</sup>

### **Diet**

Diet is a major determinant of microbiome composition. Unlike other -omics layers, diet can be directly modified, allowing researchers to study the effects of dietary intervention on host physiology. Diet data collected from human studies is usually self-reported, taking the form of food-frequency questionnaires or single-day food records.<sup>32</sup> More precise dietary measurements can be made, such as through interviews with dietary professionals, but usually at higher cost. After data collection, diet data is harmonized into a standardized form and may be converted into nutrient values prior to downstream analysis.<sup>33</sup>

### **Clinical outcomes**

With the increased use of electronic health record systems over the past several decades, clinical data has also been incorporated into multi-omic analyses. Clinical data includes medical diagnoses, lab results, doctors' notes, and sensor data such as

from glucose monitors, among other health-related information.<sup>34</sup> When combined with omics data, analysis of health record data can provide insight into the clinical consequences of -omics measurements, such as whether certain genotypes are associated with disease states.<sup>34</sup> Health record data is often incomplete and noisy, and therefore it must be harmonized to a standardized form prior to analysis.<sup>34</sup> Tong et al. include a comprehensive discussion on incorporating genomics with clinical information in studies of human health.<sup>34</sup>

### Challenges in microbiome multi-omics integration

The simultaneous analysis of many different data layers can provide greater insights into biological systems than the analysis of single layers separately.<sup>35</sup> Different datasets offer complementary views into the same biological system, elucidating relationships between different cellular processes, for example between mRNA transcript and protein levels or between particular protein structural motifs and translational pausing sites.<sup>36</sup> Integrative approaches that study the relationships between different types of biomolecules have had applications such as predicting risk of relapse in prostate cancer<sup>37</sup> and interrogating disease mechanisms in chronic kidney disease.<sup>38</sup> An extensive literature details multi-omics techniques and analysis strategies in a variety of application domains.<sup>39,40</sup> In multi-omics analyses, each -omics layer poses its own unique set of challenges. Metagenomic analyses in particular must overcome ambiguity in taxon assignments, compositionality and sparsity of data, variability of the microbiome over time, sensitivity of results to the analysis pipelines used, and a paucity of publicly available multi-omic microbiome datasets. We review these challenges below.

#### Ambiguity in taxon assignments

Microbial taxonomic labels are generally assigned imprecisely. In shotgun sequencing, abundances are inferred based on counts of short reads in sequencing experiments (usually <300bp in length) that are

aligned against multiple reference genomes to determine their taxonomic origin. Due to the immense genetic variation and diversity in the human microbiome (Tierney et al. observed that 50% of genes observed in the oral and gut metagenomes were “singletons” present in one sample but not in any others), sequences may either not match any reference genome or may match to multiple reference genomes.<sup>41</sup> Various approaches have been taken to assign sequences to taxonomies despite this ambiguity. Kraken, for example, is a popular alignment program that assigns shotgun sequencing reads to taxonomies using a maximum-score strategy.<sup>42</sup> In 16S amplicon sequencing, taxonomic classification has generally been performed by clustering sequences at an arbitrary pre-defined threshold, such as 97% or 99% sequence similarity, with the resulting taxonomic assignments known as operational taxonomic units (OTUs).<sup>43</sup> As improvements in sequencing technology have led to a decrease in sequencing error rates, taxonomies can now be assigned more precisely using methods such as amplicon sequence variants (ASVs) or zero-radius OTUs (zOTUs), which offer single-nucleotide resolution in resolving amplicons, while also allowing for the use of sophisticated error models to correct sequences that might contain errors.<sup>43–45</sup> Nonetheless, 16S sequencing approaches have been shown to exhibit lower resolution compared to shotgun sequencing.<sup>46,47</sup> Classification specificity also varies based on the particular 16S subunit that is being sequenced.<sup>48</sup> Studies should therefore be careful to ensure that their results do not reflect sequencing or other technical artifacts instead of true biological signal.

#### Compositionality

Microbiome data is compositional, in that read counts for a particular sample can only provide information about the relative abundances of one taxon relative to others instead of absolute counts.<sup>49</sup> Values in a compositional dataset represent fractions of a whole and are therefore constrained to sum to 1, lying on an *Atchinson simplex* rather than on the full Euclidean space.<sup>50</sup> To address this, methods such as the additive or centered log ratio transform have been developed to transform compositional data into a form in which it can be readily analyzed

with non-compositional analysis techniques such as linear models. Studies of the microbiome should be careful to use statistical models that account for compositionality and to avoid erroneous conclusions about the absolute abundances of taxa.

### **Sparsity**

Microbiome data is oftentimes sparse, with a high proportion of zero counts for taxa in many samples. Zero-counts may not reflect true biological signal, but may instead be an artifact of technical factors such as metagenomic inference pipeline parameters or sequencing depth.<sup>51</sup> Sparsity also limits the applicability of existing statistical models which do not explicitly account for the inflation of zero-counts. For example, dimensionality reduction methods such as PCA and PCoA may generate “horseshoe” patterns when applied on sparse data where samples do not share any taxa with each other.<sup>52</sup> The Tweedie distribution, instead of the more common Poisson or negative binomial distribution, has also been used to model zero-inflated microbial abundances.<sup>53</sup>

### **Variability**

The gut microbiome is highly variable over time and is strongly influenced by factors such as antibiotic use and diet.<sup>54</sup> A study by Vandeputte et al. found greater day-to-day variation in microbiome composition within individuals than between individuals.<sup>55</sup> Johnson et al. also showed that microbiome profiles correlate with diet, and that these correlations vary at the individual level.<sup>56</sup> Furthermore, in animal models, the microbiome is also affected by caging, bedding, food, environment, and psychological stressors.<sup>57</sup> Studies should be careful not to draw conclusions from transient measurements of the microbiome, such as by increasing sample sizes, controlling for factors such as diet, or performing repeated sampling.<sup>32</sup>

### **Small sample sizes**

While traditional genome-wide association studies have benefited in recent years by the collection of sample sizes in the millions of participants over thousands of traits,<sup>58</sup> multi-omics studies currently

have much fewer samples, usually in the tens or hundreds of participants,<sup>59</sup> with only a small subset also including microbiome data. Several repositories have been developed to address the lack of large microbiome datasets. For example, the Human Microbiome Project represents a recent effort to collect a large number of microbiome samples to facilitate large-scale computational analyses. As a part of the Human Microbiome Project, the Inflammatory Bowel Disease Multi’omics Database (IMDMDB) is a multi-omics repository specifically targeting inflammatory bowel disease.<sup>60</sup> The related Integrative Human Microbiome Project (iHMP) has also generated comprehensive integrated microbiome data in the context of preterm birth, inflammatory bowel disease, and type 2 diabetes.<sup>61</sup> In addition, the TAILORED-Treatment consortium, completed in 2018, generated a multi-omics dataset of 1,200 clinical samples containing genotype, metabolomics, and metagenomics data. However, the HoPOIT database, in which these results are stored, is only available to specific research partners and not to the public.<sup>62</sup> Despite the existence of these projects, few others also include microbiome multi-omics data. Due to the lack of integrated microbiome data in large genomic projects, researchers have turned to creative solutions. For example, Poore et al. have recently leveraged sequencing data from The Cancer Genome Atlas to infer per-sample microbial abundances from existing human WGS and transcriptomics reads.<sup>63</sup> Similarly, Dohlman et al. also used inferred microbial abundances using sequencing data associated with TCGA samples to construct the Cancer Microbiome Atlas, in a pipeline that emphasizes the decontamination of microbial reads.<sup>64</sup>

### **Sensitivity to analysis pipelines**

Bioinformatics analyses can be sensitive to pipeline parameters, providing different results when different differential abundance software, alignment pipelines, or reference databases are used.<sup>31,65,66</sup> When multiple -omics layers are processed together, this effect may be compounded further. Studies should therefore be mindful of this potential variability in results by validating their conclusions with multiple different approaches.

Furthermore, when data is combined from multiple sources, such as when performing a meta-analysis or when combining data from different cohorts of individuals, researchers should be careful to reduce pipeline-related batch effects by ensuring that samples are processed uniformly.

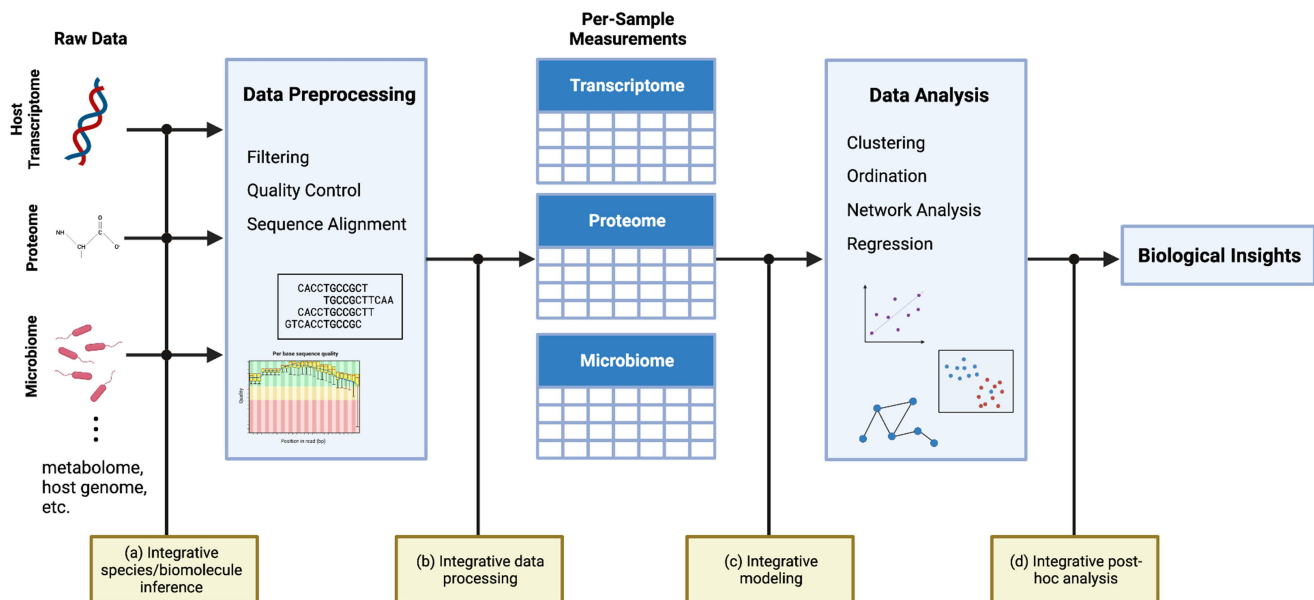
### Reliance on databases

In addition to these challenges, multi-omic bioinformatic analyses are dependent on curated databases that support the analysis of each individual data type. For example, 16S amplicon sequencing requires microbial sequences to be aligned to known rRNA sequences, which are stored in sequence databases such as SILVA.<sup>67</sup> Similarly, human metabolomics studies may rely on databases such as HMDB (Human Metabolome Database)<sup>68</sup> that provide detailed information for a multitude of metabolites; and transcriptomics studies may be reliant on pathway databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes)<sup>69</sup> that contextualize the abundances of various transcripts. The inclusion of additional layers into analysis pipelines may require the development of additional databases, which can be time-consuming to create.

Due to the difficulty in collecting data from several -omics layers over a large number of samples, studies may choose to combine data from many different projects into a single meta-analysis. Such studies must be careful to account for batch effects. Microbiome analyses are sensitive to the taxon inference and differential abundance pipelines used.<sup>65</sup> Furthermore, in low-biomass samples, technical confounders such as sample procurement strategy may also introduce effects that are greater in magnitude than the treatments themselves.<sup>70</sup>

### Current approaches to microbiome multi-omics integration

While it is clear that integrating multiple omics layers provides benefits when studying the microbiome, there is no consensus on the best way of doing so. Integration may occur at various points in the analysis pipeline, and strategies vary with the study design and the questions being asked (Figure 1). In some studies, the inference of microbial counts is itself multi-omic, with abundances being estimated from metagenomic, metatranscriptomic and metaproteomic data at the outset<sup>71</sup>



**Figure 1.** The integration of -omics layers can occur at various stages in the analysis pipeline. (a) information from multiple -omics layers can be combined to inform prediction of taxon or protein abundances. (b) multi-omics information may be combined after sequence alignment to inform per-sample measurements, such as by performing integrative batch correction. (c) per-sample measurements can be modeled together, such as in a linear or graphical model. (d) Analyses can also be performed on -omics layers individually and conclusions can be drawn from the combined analyses at the end of the study.

(Figure 1a). In others, data preprocessing steps such as batch correction are performed using information from multiple -omics layers<sup>72</sup> (Figure 1b). Others make use of computational methods that integrate data from several -omics layers simultaneously (Figure 1c). Another approach is to perform multiple separate sequencing experiments and analyze each -omics layer individually, such as in a study by Forslund et al. that investigated the effects of medications on both the gut microbiome and blood metabolome<sup>73</sup> (Figure 1d). Below we provide an overview of methods that have been used to perform microbiome analysis on multiple -omics layers, categorized based on their underlying statistical framework.

### Dimensionality reduction and clustering methods

Dimensionality reduction is often the first step in any -omics analysis because it provides a quick way to visualize the overall structure of a dataset. The most common type of dimensionality reduction, Principal Component Analysis (PCA), linearly projects a high dimensional dataset onto the data axes with the highest variance so that it can be plotted in two dimensions. Principal Coordinates Analysis (PCoA, also known as Multidimensional Scaling) is another method which produces a two-dimensional embedding of the data such that distances between data points are preserved as closely as possible. PCA or PCoA plots are usually constructed separately for each data layer prior to integrative analysis in order to identify clear patterns in the data, such as whether data points are separated between disease and control groups.<sup>74,75</sup>

Several other dimensionality reduction strategies have also been proposed, including Isomap, t-SNE, and UMAP, each performing a different transformation to embed the data onto two-dimensional space.<sup>76</sup> Dimensionality reduction methods have also been developed that operate on multiple omics datasets at the same time. Multi-omics Factor Analysis (MOFA), for example, is a linear multi-table integration method that finds a small set of numeric factors that best describe samples in a dataset.<sup>77</sup> MOFA takes as input a series of  $M$  data matrices, one for each modality over a common set of  $N$  samples, and a pre-specified number of factors

$F$ . It then decomposes each input matrix as the product of a common  $N \times F$  factor matrix and a modality-specific  $F \times D_m$  weight matrix, where  $D_m$  is the number of covariates measured in modality  $m$ . MOFA imposes regularization on the weight matrices, resulting in sparse sets of features that are associated with each factor. It was used by Garcia-Etxebarria et al. to identify genotypic, microbiome, and metabolomic factors associated with adenoma and colorectal cancer risk in a cohort of 120 individuals.<sup>78</sup> Meng et al.<sup>79</sup> include a detailed discussion about other integrative dimensionality reduction methods. Another common strategy when working with multiple data modalities is to summarize an entire data layer as a single numerical value that is then used in further analysis. Wang et al., for example, summarized multi-omic microbiome data into a single *risk score* measurement that was then used to predict disease susceptibility.<sup>80</sup> In the context of image analysis, Zhao et al. also summarized MRI scan data as a single number representing gray matter volume before integrating it with other data types.<sup>74</sup>

Clustering algorithms can also be used to identify overall patterns in a dataset. Clustering can be performed either on samples, such as when clustering patients by their metabolic markers in order to infer disease subtypes,<sup>81</sup> or on covariates, such as when identifying clusters of co-expressed genes.<sup>82</sup> In clustering, elements (either samples or covariates) are partitioned into groups, called clusters, such that elements within a cluster are more similar to each other than to elements outside of their cluster. Measures of similarity between elements may be defined in a variety of ways. Common approaches include Euclidean distance, Manhattan distance, or Bray-Curtis dissimilarity.

Clustering analyses have the potential to identify patterns of microbiome composition that have clinical significance. Clustering has been used to assign humans as having different “enterotypes” based on their intestinal microbiota composition,<sup>83</sup> which may describe alterations of the microbiome that are associated with different disease states.<sup>84</sup> Additional studies have identified clusters in the nasal microbiome associated with disease states. For example, Lehtinen et al. identified nasal microbiome clusters that predicted viral load and symptom severity in rhinovirus infection.<sup>85</sup> Similarly,

Abdel-Aziz et al. identified two distinct sputum microbiome clusters in asthma patients that were predictive of gene expression and protein abundances.<sup>86</sup>

Several approaches also support the integrative clustering of multi-omics datasets,<sup>87</sup> which allows cluster assignments to capture more complex relationships between -omics layers. This may be helpful when finding modules of co-regulated or co-occurring biomolecules or when identifying disease subtypes based on molecular signatures. iCluster, for example, is a popular matrix factorization algorithm that uses multiple -omics layers simultaneously to infer sample cluster assignments, and has been used to identify cancer subtypes based on copy number and gene expression data.<sup>88</sup> Given a series of  $n$  input data matrices and a pre-specified number of clusters  $K$ , iCluster decomposes each data matrix  $X_i$  as the product of a common cluster assignment matrix  $Z$  and  $n$  layer-specific weight matrices  $W_i$ , such that the variance explained by the cluster assignments is maximized. iCluster also imposes an L1 penalty on the weight matrices, so that each cluster is explained by a sparse set of covariates.

Several studies have used multi-omics clustering approaches to identify cancer subtypes. For example, Yuan et al. used multi-omics clustering to identify subtypes of prostate and breast cancer that were associated with patient clinical outcomes and survival.<sup>89</sup> They proposed Patient-specific Data Fusion (PSDF), a Bayesian approach that both estimates the number of disease subtypes as well as performs cluster assignment. Similarly, Chaudary et al. used a deep-learning clustering approach based on autoencoders to identify cancer subtypes that were predictive of survival outcomes.<sup>90</sup> Another approach, LRACluster, uses a maximum likelihood approach combined with K-means clustering to infer cluster assignments in continuous, binary, and count data.<sup>91</sup>

Spectral clustering is another clustering technique that has seen applications in genomic datasets. Given a dataset of size  $n$  and a pre-specified number of clusters  $k$ , spectral clustering first computes a similarity matrix that describes pairwise relationships between data points. Then, an eigendecomposition is performed on the graph Laplacian of this matrix and the first  $k$  eigenvectors are taken,

which represent inter-node relationships in a lower-dimensional subspace. Finally,  $k$ -means clustering is performed on the  $n \times k$  matrix described by the eigenvectors, which results in the final cluster assignments. Given an appropriate similarity function, spectral clustering is able to identify clusters of arbitrary shapes in high-dimensional datasets. Von Luxburg provides an in-depth explanation on spectral clustering and describes how it is equivalent to an  $n$ -cut problem on graphs.<sup>92</sup> Extensions of spectral clustering have been developed for multi-omics datasets, such as in one approach taken by Zhang et al. that finds clusters of cells in multi-modal single-cell sequencing data.<sup>93</sup>

### Correlation-based methods

Approaches to identifying relationships between quantitative covariates typically involve the computation of some pairwise measure of similarity, such as Pearson's correlation coefficient. For example, one might compute pairwise correlation coefficients between the abundances of several microbes and the expression levels of several genes. Pearson's correlation is the most common measure of similarity, but it cannot identify non-linear associations and tends to find spurious associations in compositional datasets.<sup>94</sup> Spearman correlation is an alternative method where covariates are first rank-transformed, so that each measurement is replaced with its integer-valued rank between 1 and  $N$ , where  $N$  is the number of samples, and the Pearson correlation is then computed on the ranks. Spearman correlations can find monotonic non-linear associations but still are prone to finding spurious associations in compositional data. Kendall's tau is another nonparametric measure of association between two sets of quantitative measurements that operates on ranked values from each dataset. Liu et al. include an explanation of Kendall's tau with a comparison to Spearman's correlation.<sup>95</sup> Kendall's tau has been used by Nayfach et al. to find associations between protein family abundances and clinical variables in patients with inflammatory bowel disease.<sup>96</sup>

Correlation-based methods applied to the microbiome must be aware of the risks of compositionality. Spearman and Pearson correlation coefficients



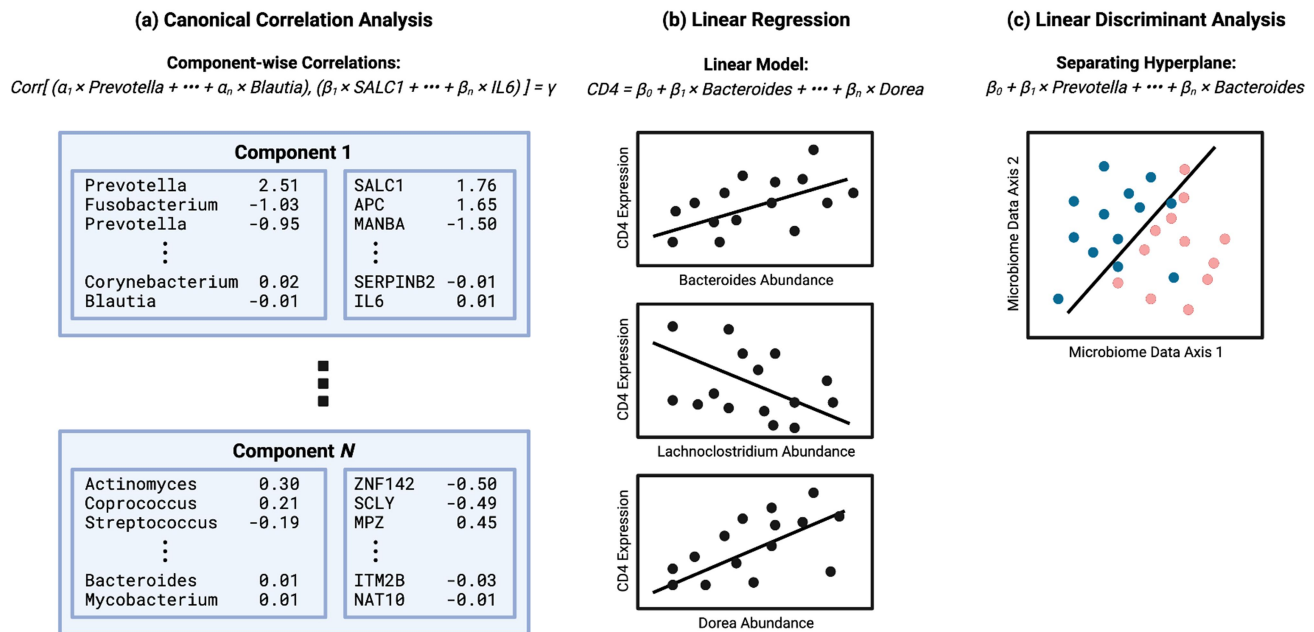
cannot be computed directly on compositional datasets without introducing spurious associations, and therefore several methods exist to transform compositional data from simplicial space into Euclidean space so that it can be analyzed with existing methods such as linear models.<sup>50</sup> One such commonly-used transformation, the centered-log-ratio (CLR) transform, replaces each measurement with the logarithm of the ratio of the measurement against the sample's geometric mean,  $m'_{ij} = \log(m_{ij}/g(m_i))$ , where  $m_{ij}$  is the  $j$ th measurement in the  $i$ th sample and  $g(m_i)$  is the geometric mean of all measurements in the sample. It is then possible to find Pearson or Spearman's correlations between pairs of CLR-transformed covariates.<sup>49</sup> Friedman et al. have also proposed SparCC, an alternative approach for finding correlations in compositional datasets, that performs better than CLR-transformed Spearman correlations on data with very few taxa but assumes sparsity of the correlation matrix, an assumption that is often useful in practice but might not always be appropriate.<sup>97</sup> Studies that make use of correlation-based methods should also be careful to not draw conclusions about causal relationships, which in general cannot be inferred from observational data.

In addition to correlations, a number of different measures of similarity have been proposed, including REBACCA,<sup>98</sup> mutual information,<sup>99</sup> and cosine similarity.<sup>100</sup> One study by Faust et al. used multiple measures of similarity at once, using an ensemble of Bray-Curtis dissimilarity, Kullback – Leibler divergence, Pearson correlation, and Spearman correlation to construct a microbial co-occurrence network.<sup>101</sup> You et al. compared six different measures of metabolite-microbe similarity and concluded that each measure of similarity is best suited for a different scenario, with Spearman correlation being the most performant overall.<sup>102</sup>

As an example of a correlation-based analysis, Wang et al. computed Pearson correlations between microbial, metabolite and host gene transcript abundances in colonic tumor and tumor-adjacent tissue and concluded that the genus *Fusobacterium* may interfere with butyrate synthesis, possibly resulting in tumorigenesis.<sup>103</sup> Spearman correlations were also used by Lloyd-Price et al. to construct a network of interactions between microbiome, host transcriptome, proteome, metabolome, and

virome that revealed disease-associated features of interest for follow-up analysis, including a highly connected but yet-unclassified microbe with genus *Subdoligranulum* as well as several acylcarnitine metabolites.<sup>60</sup> Several additional studies have also performed multi-omics analysis by finding feature-by-feature correlations.<sup>104,105</sup>

Canonical correlation analysis (CCA) is a method that finds sets of highly correlated covariates across paired datasets.<sup>106</sup> Given two data matrices representing data for the same samples, such as a table of microbe abundances and a matched table of transcript abundances, CCA finds subsets of features in each dataset that have the maximum correlation with each other (Figure 2a). The features in each component are also assigned weights so that they each represent a linear combination of the covariates in the data matrices. The components are then subtracted from the original datasets and the CCA procedure is applied again to generate several additional CCA components, each of which is orthogonal to all others. The results of CCA can sometimes be difficult to interpret, as the generated components typically include nonzero coefficients for all of the covariates. To improve the interpretability of the components generated, a variant called sparse CCA (sCCA) imposes an L1 penalty on the CCA procedure to limit the number of nonzero coefficients.<sup>107</sup> An example component from sparse CCA with an appropriate penalty parameter may include just a few microbes and their associated genes. Priya et al. used sparse CCA to characterize associations between the gut microbiota and host transcriptome in patients with colorectal cancer, inflammatory bowel disease, and irritable bowel syndrome, identifying common and disease-specific microbe-transcript interactions across all three disease types.<sup>13</sup> Hyuk Park et al. similarly used CCA to relate microbiome abundances with gene expression in gastric cancer and identified associations between Helicobacteraceae and inflammation-related genes as well as between Pasteurellaceae and Lachnospiraceae and cancer-related genes.<sup>108</sup> In addition, multivariate generalizations of CCA such as sparse multiple CCA have been developed that integrate three or more data tables simultaneously.<sup>109</sup> As an example, Galié et al. used sparse generalized CCA (SGCCA) in a study



**Figure 2.** Comparison of linear modeling strategies. (a) Canonical Correlation Analysis (CCA) finds linear combination of covariates in each –omics layer that have a maximal correlation. (b) Linear regression identifies linear relationships between a response variable, such as gene expression, and many explanatory variables, such as taxon abundances. (c) Linear discriminant analysis (LDA) finds a best separating hyperplane between two sets of data points.

of metabolic syndrome to identify four major clusters exhibiting crosstalk between gut microbiota, fecal metabolites, and host plasma metabolites.<sup>110</sup>

Procrustes analysis is another common technique to compare microbiome data and another set of measurements over a common set of samples. Procrustes analysis measures the extent to which one set of measurements can be translated, rotated, and scaled to best fit the other.<sup>111</sup> It has been frequently applied in the context of gut microbiome research, for example to find associations between the gut microbiome and diet,<sup>56</sup> and between the gut microbiome and host gene expression.<sup>13</sup>

## Regression and classification methods

Regression and classification methods attempt to predict one set of variables from another, such as disease subtype from gene transcript abundances. The variables to be predicted, such as disease subtype, are known as the response or dependent variables. The variables from which a prediction is made, such as gene expression, are known as the covariates or independent variables. Regression and classification methods are useful for identifying which features are predictive of an

outcome. For example, they may be used to identify the microbes which are most predictive of disease severity in colon cancer. Linear models represent the most common form of regression for quantitative covariates, where the response is predicted as a linear function of the covariates and is modeled to have Gaussian-distributed noise (Figure 2b).

Generalized linear models extend linear models by allowing the response to depend on the data based on a possibly nonlinear link function and to follow a non-Gaussian noise distribution. Generalized linear models were used by Tipton et al. to find associations between the oral microbiota and lab cytokine measurements.<sup>112</sup> Logistic regression is a type of generalized linear model that uses a logit link function and has been applied to find associations between gut microbiome composition and the onset of dementia.<sup>113</sup> Polynomial regression can also be used to capture nonlinear dependencies in metagenomic and multi-omics datasets.<sup>114</sup> In polynomial regression, the data table for a set of covariates is augmented with additional covariates that take the values of various powers of the original data, such as the squares and cubes of normalized microbial abundances. Linear

regression is then performed on the augmented data table.

In settings where the number of predictors is greater than the sample size, as is often the case with multi-omics datasets, it is possible to perform linear regression using penalized approaches such as Lasso,<sup>115</sup> Ridge,<sup>116</sup> or Elastic net regularization.<sup>117</sup> Regularized regression approaches simultaneously perform regression and variable selection and can limit the number of covariates in the generated model to a smaller, more interpretable subset. The HOMINID framework, for example, uses Lasso regression and stability selection to find associations between host SNPs and microbial taxa. Given datasets of host genetic variation and microbiome composition, LASSO regression is performed with host genetic variants as the model outcome and microbial taxonomic relative abundances as the covariates. Subsequently, stability selection is used to identify robust microbiome features that are associated with each host variant. It was applied on samples from the Human Microbiome Project across 15 body sites.<sup>118</sup>

A nested elastic net design was used by Ghaemi et al. to predict gestational age of pregnancy from cell-free transcriptomics, plasma and serum cytokine concentrations, metagenomics, blood mass cytometry, untargeted metabolomics, and targeted plasma proteomics.<sup>119</sup> Elastic net models were first trained to predict gestational age individually from each of the seven data modalities. These models were then combined by training a second elastic net model that predicted gestational age from the outputs of these single-modality elastic net models. This nested model identified a possible regulatory interaction between the steroid hormone pregnenolone sulfate and myeloid dendritic cells and regulatory T cells, which are known to play a critical role in the maintenance of pregnancy.

Mills et al. used linear regression to analyze metabolomic, metagenomic, metapeptidomic, metaproteomic, and host proteomic data in a cohort of 250 patients with inflammatory bowel disease.<sup>120</sup> They performed univariate linear regression to predict disease severity separately from each covariate and found that *Bacteroides* proteins, particularly proteases originating from *B. vulgatus*, were strongly associated with ulcerative colitis. They then performed a functional analysis that revealed that proteases were most

correlated with disease severity. Finally, they demonstrated *in vivo* that protease-inhibition reduced inflammation in mice inoculated with *B. vulgatus*, demonstrating a possible therapeutic target for ulcerative colitis.

Linear discriminant analysis (LDA) is a classification method that finds a best-fit hyperplane that separates labeled samples (Figure 2c). It can be used to determine the axes of variation that differ between two classes of individuals, such as those with and without a particular phenotype. As an example, Gomez-Llorente et al. performed an analysis using LDA to study the relationship between weight and intestinal microbiota composition in patients with asthma. They used sparse linear discriminant analysis (sPLS-DA) to construct three classifiers: *obese vs. non-obese*, *overweight vs. non-overweight* and *normal-weight vs. non-normal weight*.<sup>121</sup> They found that leptin, acetate, and bacteria from the order Clostridiales were predictive of *normal vs. non-normal* weight. In the same study, they also identified 12 molecular features that differentiated persistent asthma from non-persistent asthma, such as creatinine and citrate concentrations.

Random forests are a commonly used classification method in which many decision trees, each individually weak and trained on a random subset of the features, are used to construct a stronger classifier that can discriminate between outcomes of interest such as disease state. Silveira et al., in a study including metagenomics, metabolomics and fluorescence microscopy, used random forests to study the role of microbes in the progression of cystic fibrosis. They identified facultative anaerobes such as *Streptococcus* as being keystone bacteria in the cystic fibrosis microbiome.<sup>75</sup> Zeybel et al. also used random forests to identify genomic features associated with hepatic steatosis in a cohort of 78 patients.<sup>122</sup> Zhao et al. similarly used random forests and linear discriminant analysis in a cohort of 50 patients to study the gut microbiome in major depressive disorders using metagenome, metabolome, inflammatory factor and MRI data.<sup>74</sup>

## Network methods

An ultimate goal of multi-omics studies is to generate a comprehensive map of interactions between

the various molecules and organisms within a system.<sup>123</sup> Networks, also known as graphs, provide an intuitive approach for visualizing and studying these interactions. Unlike regression-based methods that find associations between two groups of variables at a time, network-based methods aim to model all of the interactions between entities simultaneously. Such methods have been utilized to study co-occurrence relationships between microbial taxa,<sup>124</sup> co-abundance relationships between metabolites,<sup>125</sup> and co-expression relationships between genes.<sup>126</sup>

Graphs are specified by a set of *nodes*, which usually refer to covariates such as gene expression values, microbial abundances, or disease states; and a set of *edges*, numerical values that connect pairs of nodes and which are usually a measure of similarity such as Pearson correlation or mutual information. A graph may be generated from a dataset by first finding correlations between all pairs of covariates, next identifying which correlations are significant, and then finally connecting pairs of nodes that are significantly correlated with each other.<sup>83</sup> After a graph is constructed, various analytics can be performed on it. For example, a “clique” analysis can identify highly interconnected nodes, while a “connectivity” analysis can identify distinct groups of nodes.<sup>127</sup> Layeghifard et al. provide a detailed review of general network approaches applied to microbiome datasets.<sup>128</sup> In some applications, known pathway information can also be integrated with graph inference.<sup>129</sup> Liu et al. provide a comprehensive treatment of network methods for the analysis of microbiome data.<sup>130</sup>

Cantoni et al. used correlation-based networks to study the gut microbiome, blood immune cell concentrations, and circulating metabolites in patients with multiple sclerosis.<sup>131</sup> They constructed a pairwise association network separately for MS patients and controls, connecting covariates if they had a pairwise correlation coefficient greater than 0.7. They found that memory Th1 cells were correlated with many metabolites in controls but not in MS patients; they also identified a significant pathway linking *B. thetaiotaomicron*, Th17 cells, S-adenosylmethionine (associated with Th17 cell activation), and meat consumption.

Pfalzer et al. used a network approach to study the interactions between transcriptome,

metabolome, and microbiome in obese mouse models of colon cancer.<sup>132</sup> Three groups of mice were either fed a low-fat (LF) diet, made obese with a high-fat (HF) diet, or made obese with a leptin receptor mutation (DbDb). Differential gene expression analysis was performed twice, between HF and LF mice and DbDb and LF mice, and a co-expression network was constructed individually for both sets of differentially expressed genes. Gene modules (sets of correlated genes) were identified from both networks and correlations were then computed between these gene modules and microbial and metabolic abundances using the WCGNA R library.<sup>133</sup> Overall, the study implicated Akt-signaling genes, microbial genera including *Clostridium* and *Sarcina*, and adenosine concentration as being associated with tumor burden.

Bayesian networks are a promising approach to studying interactions between covariates and diseases.<sup>134</sup> Bayesian networks model the entire joint distribution of covariates as a directed graph, specifying the dependence relationships of variables to each other and allowing for the inference of the value of one variable given several others. While Bayesian network inference is computationally intractable for large numbers of covariates, approximation-based techniques have had success in inferring Bayesian network structure from measured data. Hill-climbing, for example, is a greedy approach where edges are sequentially added to the Bayesian network so as to maximize a measure of fit, such as the Bayesian Information Criterion, at each step. Su et al. present a treatment of Bayesian networks in the context of identifying disease-related genes.<sup>135</sup> Krishnan et al. used Bayesian networks to study multi-omics data in nonalcoholic fatty liver disease (NAFLD).<sup>136</sup> They constructed a Bayesian network describing gene-gene interactions and their interactions with pathophysiological conditions and identified potential regulatory genes associated with NAFLD.

### Spatial and temporal multi-omics

Advances in spatial analysis have enabled the integration of multi-omics data along with spatial information about the microbiome. For example, Shi et al. have demonstrated a novel technique, HiPR-FISH, which is able to produce spatial maps

of microbial taxa within the gut and may be of interest in animal studies of the microbiome.<sup>137</sup> In another application of spatial multi-omics, Garg et al. developed a method to integrate metabolomic, 16S metagenomic, and spatial information to study the organization of the lung microbiome in cystic fibrosis patients.<sup>138</sup> A lung was procured from a recently deceased CF patient and sectioned into 86 segments from which metabolomic and metagenomic data was collected. Spatial maps of the multi-omics data were generated which demonstrated that microbial composition varies within the lung and that this variation may be related to the spatial variance of disease severity.

When multi-omics measurements are taken from samples at many different time points, it becomes possible to learn about the temporal shifts in microbial interactions with the various biomolecules in a system. As gut microbial composition is highly variable in the short term and fluctuates with factors such as diet and flare status in patients with IBS or IBD, temporal analyses allow for the explicit modeling of microbiome variability over time.<sup>139,140</sup> One study by Mihindukulasuriya et al. studied the temporal dynamics in the gut virome in the context of diet, bacterial microbiome, metabolome, and host genome and transcriptome. For each of 50 study participants who were either healthy or diagnosed with IBS-C or IBS-D, two consecutive stool samples were sequenced for viral DNA. For a subset of 28 participants, colonic transcriptomics data was also collected. Their results indicated that the viral microbiome was stable in stool samples among both IBS and healthy individuals, despite the high variability of bacterial composition. They also identified immune-related genes associated with the virome and disease-specific associations with specific phage populations, indicating that the virome might play a role in regulating bacterial composition.<sup>141</sup> In another time-series analysis, Ruiz-Perez et al. fit a dynamic Bayesian network to model the time evolution of multi-omics microbiome data in inflammatory bowel disease, incorporating prior knowledge about multi-omics interactions to constrain the model to biologically plausible interactions. The Bayesian network successfully predicted taxon abundances at future time points. It also inferred

several metabolite-to-microbiome relationships which were validated experimentally.<sup>142</sup>

Temporal microbiome analyses have also been conducted in wastewater systems to obtain epidemiological insights into the prevalence of pathogenic agents in a community and to study the efficacy of sanitation practices.<sup>143</sup> Herold et al. employed a multi-omics time-series analysis to investigate the effect of disturbances on microbial communities in wastewater sludge. In order to identify the niches of microbial populations in wastewater, they collected metagenomic, metatranscriptomic, metaproteomic and metametabolomic data weekly over the course of 14 months. Multidimensional Scaling (MDS) was used to identify bacterial niches. For each of several functional categories (for example, “nucleic acid metabolism”), metatranscriptomic abundances were correlated with metaproteomics levels. Four functionally distinct niches were identified, each with different responses to environmental changes, suggesting that future research directions may involve untangling these distinct subpopulations.<sup>144</sup>

### Multi-omic taxon and biomolecule inference

The previous methods have used multi-omics data to study the interactions between different biomolecules. However, multiple -omics datasets can also be used at an earlier stage in the analysis pipeline in order to improve inference of the biomolecules themselves. gNOMO provides one such technique.<sup>145</sup> Protein identification is commonly achieved through mass spectrometry, a technique that typically allows the quantification of only a subset of the entire set of proteins in a sample.<sup>146</sup> gNOMO uses metatranscriptomics and metagenomics information within the sample to predict the set of proteins that may be encoded by bacteria and the host using a program called Prodigal. This set of likely proteins is used as the mass spectrum lookup database for mass spectrometry.

Other approaches have combined metagenomic and metatranscriptomic information to improve the inference of microbial abundances. For example, Heintz-Buschart et al. observed that a significant proportion of reads sequenced with shotgun metagenomic sequencing did not map to

a fully-sequenced reference genome.<sup>71</sup> In order to improve taxon inference, they performed co-assembly with metagenomic and metatranscriptomics reads, which resulted in longer contig lengths and higher read usage. Furthermore, they also inferred metaproteomic abundances using a protein search database informed by the observed metagenomic and metatranscriptomic sequences.

### Metabolic modeling

Microbiome metabolic modeling is an approach that constructs a mathematical representation of the entire set of biochemical reactions that occur within a microbial ecosystem. Metabolic models allow researchers to reason about inter-microbe and microbe-environment interactions. Metabolic models can be used to predict steady-state microbial abundances and the rates of metabolite consumption and secretion by bacteria in a system.<sup>147</sup> Magnúsdóttir et al. developed the AGORA (Assembly of Gut Organisms through Reconstruction and Analysis) resource which contains genome-scale metabolic constructions for 773 human gut bacteria.<sup>148</sup> AGORA was initially applied to identify a growth medium for *Bacteroides caccae*. In another application, a metabolic model using AGORA was used to identify differences in microbial metabolism between patients with IBD and healthy controls.<sup>149</sup>

Metabolic models support the integration of multi-omics data by imposing constraints based on metagenomic, metabolomic, metaproteomic, or diet information.<sup>149</sup> Yizhak et al., for example, developed IOMA (integrative omics-metabolic analysis), a method that integrates metabolic reconstructions with matched proteomics and metabolomics measurements in order to find a set of metabolic fluxes that is consistent with the known metabolic constructions as well as with measured data.<sup>150</sup>

### Conclusions and future directions

Integrative analysis has the potential to provide insights into disease biology by illuminating the complex interplay between the host, microbiome, and the various biomolecules through which they

interact. The integration of multiple -omics layers can elucidate interactions that may not be apparent when considering data layers individually, which improves the discovery of disease mechanisms.<sup>120</sup>

An emerging application of multi-omics analysis is in precision medicine, where measurements from multiple -omics layers are used to inform treatment decisions, such that care is targeted toward the particular physiology of the patient. Due to the multifactorial role of the microbiota, such as in the production of bioactive compounds and the metabolism of pharmaceutical drugs, the microbiome provides a promising target for precision medicine. For example, it may be beneficial to adjust medications or dosages based on a patient's microbiome composition or on other molecular phenotypes.<sup>151</sup> Therapies may also use multi-omics information to target the microbiome itself, such as by selectively modulating bacterial abundances based on host physiology.<sup>152</sup>

While a variety of approaches have been developed to support multi-omics integration, there is a current lack of standardization among them, which makes it difficult to interpret whether the results of multi-omic studies capture true signal or pipeline-related artifacts. A set of best practices should be established for integrating data layers together, which would help clarify the approaches that are best suited for each experimental design and improve the ability to compare results between studies. Standardization is especially relevant for microbiome analyses, which are sensitive to the pipeline parameters and bioinformatics software used.<sup>65</sup> Another important step in standardization is the development of readily accessible software packages that support multi-omics microbiome integration, as few currently exist.<sup>153</sup> Several software packages support the integration of multiple -omics layers in general.<sup>154</sup> However, these methods do not specifically address the challenges relevant to microbiome analysis, and it is necessary to benchmark how well they perform on microbiome datasets.

Another major difficulty in conducting a multi-omics study is the high cost of sample collection and data generation, which limits the ability of individual studies to detect associations with small effect sizes. Genetics research has benefited by the creation

of large repositories, such as ENCODE and TCGA, that aggregate multi-omic genotypic and phenotypic data across different studies, and which researchers can use to perform large-scale computational analyses.<sup>155,156</sup> It would be beneficial for such repositories to be developed and expanded in the context of microbiome multi-omics, with an emphasis on taking into account differences in pipeline parameters and batch effects between studies.

## Acknowledgments

We thank Rich Abdill for insightful comments on this manuscript.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

The work was supported by the National Institutes of Health [R35-GM128716 to RB].

## ORCID

Ashwin Chetty  <http://orcid.org/0000-0002-1395-3930>

Ran Blekhman  <http://orcid.org/0000-0003-3218-613X>

## References

- Jandhyala SM, Talukdar R, Subramanyam C, Vuyyuru H, Sasikala M, Nageshwar Reddy D. Role of the normal gut microbiota. *World J Gastroenterol* [Internet]. 2015;21(29):8787–8803. doi:10.3748/wjg.v21.i29.8787.
- Rebersek M. Gut microbiome and its role in colorectal cancer. *BMC Cancer* [Internet]. 2021;21(1):1325. doi:10.1186/s12885-021-09054-2.
- Santana PT, Rosas SLB, Ribeiro BE, Marinho Y, de Souza HSP. Dysbiosis in inflammatory bowel disease: pathogenic role and potential therapeutic targets. *Int J Mol Sci Int*. 2022;23(7):3464. doi:10.3390/ijms23073464.
- Breton J, Galmiche M, Déchelotte P. Dysbiotic gut bacteria in obesity: an overview of the metabolic mechanisms and therapeutic perspectives of next-generation probiotics. *Microorganisms*. 2022;10(2):452. doi:10.3390/microorganisms10020452.
- Cheung SG, Goldenthal AR, Uhlemann A-C, Mann JJ, Miller JM, Sublette ME. Systematic review of gut

- microbiota and major depression. *Front Psychiatry* [Internet]. 2019;10:34. doi:10.3389/fpsy.2019.00034.
- Hiergeist A, Gläsner J, Reischl U, Gessner A. Analyses of Intestinal Microbiota: Culture versus Sequencing: Figure 1. *ILAR J* [Internet]. 2015;56(2):228–240. doi:10.1093/ilar/ilv017.
- Lau JT, Whelan FJ, Herath I, Lee CH, Collins SM, Bercik P, Surette MG. Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Med*[Internet]. 2016;8(1):72. doi:10.1186/s13073-016-0327-7.
- Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. A new genomic blueprint of the human gut microbiota. *Nature* [Internet]. 2019;568(7753):499–504. doi:10.1038/s41586-019-0965-1.
- Cao X, Dong A, Kang G, Wang X, Duan L, Hou H, Zhao T, Wu S, Liu X, Huang H, et al. Modeling spatial interaction networks of the gut microbiota. *Gut Microbes* [Internet]. 2022;14(1):2106103. doi:10.1080/19490976.2022.2106103.
- Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL. Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* [Internet]. 2019;570(7762):462–467. doi:10.1038/s41586-019-1291-3.
- Richards AL, Muehlbauer AL, Alazizi A, Burns MB, Findley A, Messina F, Gould TJ, Cascardo C, Pique-Regi R, Blekhman R, et al. Gut microbiota has a widespread and modifiable effect on host gene regulation. *mSystems*. 2019;4(5). doi: <http://dx.doi.org/10.1128/mSystems.00323-18>.
- Dayama G, Priya S, Niccum DE, Khoruts A, Blekhman R. Interactions between the gut microbiome and host gene regulation in cystic fibrosis. *Genome Med* [Internet]. 2020;12(1):12. doi:10.1186/s13073-020-0710-2.
- Priya S, Burns MB, Ward T, Mars RAT, Adamowicz B, Lock EF, Kashyap PC, Knights D, Blekhman R. Identification of shared and disease-specific host gene–microbiome associations across human diseases using multi-omic integration. *Nat Microbiol* [Internet]. 2022;7(6):780–795. doi:10.1038/s41564-022-01121-z.
- Nichols RG, Davenport ER. The relationship between the gut microbiome and host gene expression: a review. *Hum Genet* [Internet]. 2021;140(5):747–760. doi:10.1007/s00439-020-02237-0.
- Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci* [Internet]. 2014;5:209. doi:10.3389/fpls.2014.00209.
- Thielemann N, Herz M, Kurzai O, Martin R. Analyzing the human gut mycobiome - a short guide for beginners. *Comput Struct Biotechnol J* [Internet]. 2022;20:608–614. doi:10.1016/j.csbj.2022.01.008.
- Bai G-H, Lin S-C, Hsu Y-H, Chen S-Y. The human virome: viral metagenomics, relations with human

- diseases, and therapeutic applications. *Viruses* [Internet]. 2022;14(2):278. doi:10.3390/v14020278.
18. van Tilburg Bernardes E, Pettersen VK, Gutierrez MW, Laforest-Lapointe I, Jendzjowsky NG, Cavin J-B, Vicentini FA, Keenan CM, Ramay HR, Samara J, et al. Intestinal fungi are causally implicated in microbiome assembly and immune development in mice. *Nat Commun* [Internet]. 2020;11(1):2577. doi:10.1038/s41467-020-16431-1.
  19. Heeger F, Wurzbacher C, Bourne EC, Mazzoni CJ, Monaghan MT, Yu D. Combining the 5.8S and ITS2 to improve classification of fungi. *Methods Ecol Evol* [Internet]. 2019;10(10):1702–1711. <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13266>.
  20. Houshyar Y, Massimino L, Lamparelli LA, Danese S, Ungaro F. Going beyond bacteria: uncovering the role of archaeome and mycobioime in inflammatory bowel disease. *Front Physiol* [Internet]. 2021;12:783295. doi:10.3389/fphys.2021.783295.
  21. Gaci N, Borrel G, Tottey W, O'Toole PW, Brugère J-F. Archaea and the human gut: new beginning of an old story. *World J Gastroenterol* [Internet]. 2014;20(43):16062–16078. doi:10.3748/wjg.v20.i43.16062.
  22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* [Internet]. 2010;26(1):139–140. doi:http://dx.doi.org/10.1093/bioinformatics/btp616.
  23. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* [Internet]. 2004;5(10):R80. doi:10.1186/gb-2004-5-10-r80.
  24. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, et al. Best practices for analysing microbiomes. *Nat Rev Microbiol* [Internet]. 2018;16(7):410–422. doi:10.1038/s41579-018-0029-9.
  25. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* [Internet]. 2016;17(1):13. doi:10.1186/s13059-016-0881-8.
  26. Shakya M, Lo C-C, Chain PSG. Advances and challenges in metatranscriptomic analysis. *Front Genet* [Internet]. 2019;10:904. doi:10.3389/fgene.2019.00904.
  27. Grieneisen L, Dasari M, Gould TJ, Björk JR, Grenier J-C, Yotova V, Jansen D, Gottel N, Gordon JB, Learn NH, et al. Gut microbiome heritability is nearly universal but environmentally contingent. *Science* [Internet]. 2021;373(6551):181–186. doi:10.1126/science.aba5483.
  28. Kockum I, Huang J, Stridh P. Overview of genotyping technologies and methods. *Curr Protoc* [Internet]. 2023;3(4):e727. doi:10.1002/cpz1.727.
  29. Goodrich JK, Davenport ER, Clark AG, Ley RE. The relationship between the human genome and microbiome comes into view. *Annu Rev Genet* [Internet]. 2017;51(1):413–433. doi:10.1146/annurev-genet-110711-155532.
  30. Chong J, Xia J. Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites* [Internet]. 2017;7(4):62. doi:10.3390/metabo7040062.
  31. Peters DL, Wang W, Zhang X, Ning Z, Mayne J, Figeys D. Metaproteomic and Metabolomic Approaches for Characterizing the Gut Microbiome. *Proteomics* [Internet]. 2019;19(16):e1800363. doi:10.1002/pmic.201800363.
  32. Johnson AJ, Zheng JJ, Kang JW, Saboe A, Knights D, Zivkovic AM. A Guide to Diet-Microbiome Study Design. *Front Nutr* [Internet]. 2020;7:79. doi:10.3389/fnut.2020.00079.
  33. Marconi S, Durazzo A, Camilli E, Lisciani S, Gabrielli P, Aguzzi A, Gambelli L, Lucarini M, Marletta L. Food composition databases: considerations about complex food matrices. *Foods* [Internet]. 2018;7(1):2. doi:10.3390/foods7010002.
  34. Tong L, Wu H, Wang MD, Wang G. Chapter one - introduction of medical genomics and clinical informatics integration for p-health care. In: Teplow D, editor. *Progress in molecular biology and translational science*. Academic Press; 2022. pp. 1–37. [Internet]. <https://www.sciencedirect.com/science/article/pii/S187711732200062X>.
  35. Sun YV, Hu Y-J. Chapter three - integrative analysis of multi-omics data for discovery and functional studies of complex human diseases [internet]. In: Friedmann T, Dunlap J Goodwin Seditors. *Advances in genetics*. Academic Press; 2016. pp. 147–190 <https://www.sciencedirect.com/science/article/pii/S0065266015000516>
  36. Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, Lerman JA, Lechner A, Sastry A, Bordbar A, et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nat Commun* [Internet]. 2016;7(1):13091. doi:10.1038/ncomms13091.
  37. Wei Z, Han D, Zhang C, Wang S, Liu J, Chao F, Song Z, Chen G. Deep learning-based multi-omics integration robustly predicts relapse in prostate cancer. *Front Oncol* [Internet]. 2022;12:893424. doi:10.3389/fonc.2022.893424.
  38. Eddy S, Mariani LH, Kretzler M. Integrated multi-omics approaches to improve classification of chronic kidney disease. *Nat Rev Nephrol* [Internet]. 2020;16(11):657–668. doi:10.1038/s41581-020-0286-5.
  39. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* [Internet]. 2017;18(1):83. doi:10.1186/s13059-017-1215-1.
  40. 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 [Internet]. 2021;19:3735–3746. doi:10.1016/j.csbj.2021.06.030.
41. Tierney BT, Yang Z, Lubber JM, Beaudin M, Wibowo MC, Baek C, Mehlenbacher E, Patel CJ, Kostic AD. The landscape of genetic content in the gut and oral human microbiome. *Cell Host Microbe* [Internet]. 2019;26(2):283–95.e8. doi:10.1016/j.chom.2019.07.008.
  42. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* [Internet]. 2014;15(3):R46. doi:10.1186/gb-2014-15-3-r46.
  43. Chiarello M, McCauley M, Villéger S, Jackson CR, Moreno-Hagelsieb G. Ranking the biases: the choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. *PLoS One* [Internet]. 2022;17(2):e0264443. doi:10.1371/journal.pone.0264443.
  44. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*. 2017;2(2). doi: 10.1128/mSystems.00191-16.
  45. Antich A, Palacin C, Wangenstein OS, Turon X. To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics* [Internet]. 2021;22(1):177. doi:10.1186/s12859-021-04115-6.
  46. Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. Quantitative assessment of Shotgun Metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. *OMICS* [Internet]. 2018;22(4):248–254. doi:10.1089/omi.2018.0013.
  47. Stothart MR, McLoughlin PD, Poissant J. Shallow shotgun sequencing of the microbiome recapitulates 16S amplicon results and provides functional insights. *Mol Ecol Resour* [Internet]. 2023;23(3):549–564. doi:10.1111/1755-0998.13713.
  48. Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods* [Internet]. 2007;69(2):330–339. doi:10.1016/j.mimet.2007.02.005.
  49. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* [Internet]. 2017;8:2224. doi:10.3389/fmicb.2017.02224.
  50. Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. It's all relative: analyzing microbiome data as compositions. *Ann Epidemiol* [Internet]. 2016;26(5):322–329. doi:10.1016/j.annepidem.2016.03.003.
  51. Lutz KC, Jiang S, Neugent ML, De Nisco NJ, Zhan X, Li Q. A survey of statistical methods for microbiome data analysis. *Front Appl Math stat* [Internet]. 2022;8. doi:10.3389/fams.2022.884810.
  52. Morton JT, Toran L, Edlund A, Metcalf JL, Lauber C, Knight R, Jansson JK. Uncovering the horseshoe effect in microbial analyses. *mSystems*. 2017;2(1). doi: 10.1128/mSystems.00166-16.
  53. New FN, Baer BR, Clark AG, Wells MT, Brito IL. Collective effects of human genomic variation on microbiome function. *Sci Rep* [Internet]. 2022;12(1):3839. doi:10.1038/s41598-022-07632-3.
  54. Dudek-Wicher RK, Junka A, Bartoszewicz M. The influence of antibiotics and dietary components on gut microbiota. *Prz Gastroenterol* [Internet]. 2018;13(2):85–92. doi:10.5114/pg.2018.76005.
  55. Vandeputte D, De Commer L, Tito RY, Kathagen G, Sabino J, Vermeire S, Faust K, Raes J. Temporal variability in quantitative human gut microbiome profiles and implications for clinical research. *Nat Commun* [Internet]. 2021;12(1):6740. doi:10.1038/s41467-021-27098-7.
  56. Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, Kim AD, Shmagel AK, Syed AN, Microbiome Class Students P, et al. Daily sampling reveals personalized diet-microbiome associations in humans. *Cell Host Microbe* [Internet]. 2019;25(6):789–802.e5. doi:10.1016/j.chom.2019.05.005.
  57. Wang Y, Lêcao K-A. Managing batch effects in microbiome data. *Brief Bioinform* [Internet]. 2020;21(6):1954–1970. doi:10.1093/bib/bbz105.
  58. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. *Nature Reviews Methods Primers* [Internet]. 2021 [[cited 2023 May 31]];1(1):1–21. <https://www.nature.com/articles/s43586-021-00056-9>.
  59. Gomez-Cabrero D, Tarazona S, Ferreirós-Vidal I, Ramirez RN, Company C, Schmidt A, Reijmers T, von S PV, Marabita F, Rodríguez-Ubrea J, et al. Stategra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci Data* [Internet]. 2019;6(1):256. doi:10.1038/s41597-019-0202-7.
  60. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* [Internet]. 2019;569(7758):655–662. doi:10.1038/s41586-019-1237-9.
  61. Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* [Internet]. 2014;16(3):276–289. doi:10.1016/j.chom.2014.08.014.
  62. van Houten CB, Oved K, Eden E, Cohen A, Engelhard D, Boers S, Kraaij R, Karlsson R, Fernandez D, Gonzalez E, et al. Observational multi-centre, prospective study to characterize novel pathogen-and host-related factors in hospitalized patients with lower respiratory tract infections and/or sepsis - the “TAILORED-Treatment” study.

- BMC Infect Dis [Internet]. 2018;18:377. doi:10.1186/s12879-018-3300-9.
63. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolk T, Janssen S, Metcalf J, Song SJ, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* [Internet]. 2020;579(7800):567–574. doi:10.1038/s41586-020-2095-1.
  64. Dohlman AB, Arguijo Mendoza D, Ding S, Gao M, Dressman H, Iliev ID, Lipkin SM, Shen X. The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* [Internet]. 2021;29(2):281–98.e5. doi:10.1016/j.chom.2020.12.001.
  65. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, Jones CMA, Wright RJ, Dhanani AS, Comeau AM, et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat Commun* [Internet]. 2022;13(1):342. doi:10.1038/s41467-022-28034-z.
  66. Arora S, Pattwell SS, Holland EC, Bolouri H. Variability in estimated gene expression among commonly used RNA-seq pipelines. *Sci Rep* [Internet]. 2020;10(1):2734. doi:10.1038/s41598-020-59516-z.
  67. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* [Internet]. 2013;41(D1):D590–6. doi:10.1093/nar/gks1219.
  68. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, et al. HMDB: the human metabolome database. *Nucleic Acids Res* [Internet]. 2007;35(Database):D521–6. doi:10.1093/nar/gkl923.
  69. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa MK. Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* [Internet]. 1999;27(1):29–34. doi:10.1093/nar/27.1.29.
  70. Debelius J, Song SJ, Vazquez-Baeza Y, Xu ZZ, Gonzalez A, Knight R. Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biol* [Internet]. 2016;17(1):217. doi:10.1186/s13059-016-1086-x.
  71. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* [Internet]. 2016;2(1):16180. doi:10.1038/nmicrobiol.2016.180.
  72. Ugidos M, Nueda MJ, Prats-Montalbán JM, Ferrer A, Conesa A, Tarazona S, Birol I. MultiBaC: an R package to remove batch effects in multi-omic experiments. *Bioinformatics* [Internet]. 2022;38(9):2657–2658. doi:10.1093/bioinformatics/btac132.
  73. Forslund SK, Chakaroun R, Zimmermann-Kogadeeva M, Markó L, Aron-Wisniewsky J, Nielsen T, Moitinho-Silva L, Schmidt TSB, Falony G, Vieira-Silva S, et al. Combinatorial, additive and dose-dependent drug–microbiome associations. *Nature* [Internet]. 2021;600(7889):500–505. doi:10.1038/s41586-021-04177-9.
  74. Zhao H, Jin K, Jiang C, Pan F, Wu J, Luan H, Zhao Z, Chen J, Mou T, Wang Z, et al. A pilot exploration of multi-omics research of gut microbiome in major depressive disorders. *Transl Psychiatry* [Internet]. 2022;12(1):8. doi:10.1038/s41398-021-01769-x.
  75. Silveira CB, Cobián-Güemes AG, Uranga C, Baker JL, Edlund A, Rohwer F, Conrad D. Multi-omics study of keystone species in a cystic fibrosis microbiome. *Int J Mol Sci Int*. 2021;22(21):12050. doi:10.3390/ijms222112050.
  76. Nanga S, Bawah AT, Acquaye BA, Billa M-I, Baeta FD, Odai NA, Obeng SK, Nsiah AD. Review of dimension reduction methods. *J Data Anal Inf Process* [Internet]. 2021;9(03):189–231. <http://www.scirp.org/Journal/Paperabs.aspx?paperid=111638>. doi:10.4236/jdaip.2021.93013.
  77. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* [Internet]. 2018;14(6):e8124. doi:10.15252/msb.20178124.
  78. Garcia-Etxebarria K, Clos-Garcia M, Telleria O, Nafria B, Alonso C, Iruarrizaga-Lejarreta M, Franke A, Crespo A, Iglesias A, Cubiella J, et al. Interplay between genome, metabolome and microbiome in colorectal cancer. *Cancers* [Internet]. 2021;13(24):6216. doi:10.3390/cancers13246216.
  79. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* [Internet]. 2016;17(4):628–641. doi:10.1093/bib/bbv108.
  80. Wang C, Segal LN, Hu J, Zhou B, Hayes RB, Ahn J, Li H. Microbial risk score for capturing microbial characteristics, integrating multi-omics data, and predicting disease risk. *Microbiome* [Internet]. 2022;10(1):121. doi:10.1186/s40168-022-01310-2.
  81. Pfeifer B, Schimek MG. A hierarchical clustering and data fusion approach for disease subtype discovery. *J Biomed Inform* [Internet]. 2021;113:103636. doi:10.1016/j.jbi.2020.103636.
  82. Abu-Jamous B, Kelly S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biol* [Internet]. 2018;19(1):172. doi:10.1186/s13059-018-1536-8.
  83. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, et al. Enterotypes of the human gut microbiome. *Nature* [Internet]. 2011;473(7346):174–180. doi:10.1038/nature09944.
  84. Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, de Vos WM, Ehrlich SD, Fraser CM, Hattori M, et al. Enterotypes in the

- landscape of gut microbial community composition. *Nat Microbiol* [Internet]. 2018;3(1):8–16. doi:10.1038/s41564-017-0072-8.
85. Lehtinen MJ, Hibberd AA, Männikkö S, Yeung N, Kauko T, Forssten S, Lehtoranta L, Lahtinen SJ, Stahl B, Lyra A, et al. Nasal microbiota clusters associate with inflammatory response, viral load, and symptom severity in experimental rhinovirus challenge. *Sci Rep* [Internet]. 2018;8(1):11411. doi:10.1038/s41598-018-29793-w.
  86. Abdel-Aziz MI, Vijverberg SJH, Neerincx AH, Brinkman P, Wagener AH, Riley JH, Sousa AR, Bates S, Wagers SS, De Meulder B, et al. A multi-omics approach to delineate sputum microbiome-associated asthma inflammatory phenotypes. *Eur Respir J Internet*. 2022;59(1):2102603. doi:10.1183/13993003.02603-2021.
  87. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* [Internet]. 2018;46(20):10546–10562. doi:10.1093/nar/gky889.
  88. Shen R, Olshen AB, 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* [Internet]. 2009;25(22):2906–2912. doi:10.1093/bioinformatics/btp543.
  89. Yuan Y, Savage RS, Markowitz F, Markel S. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput Biol* [Internet]. 2011;7(10):e1002227. doi:10.1371/journal.pcbi.1002227.
  90. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* [Internet]. 2018;24(6):1248–1259. doi:10.1158/1078-0432.CCR-17-0853.
  91. Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics* [Internet]. 2015;16(1):1022. doi:10.1186/s12864-015-2223-8.
  92. von Luxburg U. A tutorial on spectral clustering. *Stat Comput* [Internet]. 2007;17(4):395–416. doi:10.1007/s11222-007-9033-z.
  93. Zhang S, Leistico JR, Cho RJ, Cheng JB, Song JS, Martelli PL. Spectral clustering of single-cell multi-omics data on multilayer graphs. *Bioinformatics* [Internet]. 2022;38(14):3600–3608. doi:10.1093/bioinformatics/btac378.
  94. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J, Dunbrack RL Jr. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol* [Internet]. 2015;11(3):e1004075. doi:10.1371/journal.pcbi.1004075.
  95. Liu J, Tang W, Chen G, Lu Y, Feng C, Tu XM. Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Arch Psychiatry* [Internet]. 2016;28:115–120. doi:10.11919/j.issn.1002-0829.216045.
  96. Nayfach S, Bradley PH, Wyman SK, Laurent TJ, Williams A, Eisen JA, Pollard KS, Sharpton TJ, Guigo R. Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLoS Comput Biol* [Internet]. 2015;11(11):e1004573. doi:10.1371/journal.pcbi.1004573.
  97. Friedman J, Alm EJ, von Mering C. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* [Internet]. 2012;8(9):e1002687. doi:10.1371/journal.pcbi.1002687.
  98. Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* [Internet]. 2015;31(20):3322–3329. doi:10.1093/bioinformatics/btv364.
  99. Tackmann J, Matias Rodrigues JF, von Mering C. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Syst* [Internet]. 2019;9(3):286–96.e8. doi:10.1016/j.cels.2019.08.002.
  100. Jackson MA, Verdi S, Maxan M-E, Shin CM, Zierer J, Bowyer RCE, Martin T, Williams FMK, Menni C, Bell JT, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat Commun* [Internet]. 2018;9(1):2655. doi:10.1038/s41467-018-05184-7.
  101. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C, Ouzounis CA. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* [Internet]. 2012;8(7):e1002606. doi:10.1371/journal.pcbi.1002606.
  102. You Y, Liang D, Wei R, Li M, Li Y, Wang J, Wang X, Zheng X, Jia W, Chen T. Evaluation of metabolite-microbe correlation detection methods. *Anal Biochem* [Internet]. 2019;567:106–111. doi:10.1016/j.ab.2018.12.008.
  103. Wang Q, Ye J, Fang D, Lv L, Wu W, Shi D, Li Y, Yang L, Bian X, Wu J, et al. Multi-omic profiling reveals associations between the gut mucosal microbiome, the metabolome, and host DNA methylation associated gene expression in patients with colorectal cancer. *BMC Microbiol* [Internet]. 2020;20(S1):83. doi:10.1186/s12866-020-01762-2.
  104. Abdel-Aziz MI, Kermani NZ, Neerincx AH, Vijverberg SJH, Guo Y, Howarth P, Dahlen S-E, Djukanovic R, Sterk PJ, Kraneveld AD, et al.

- Association of endopeptidases, involved in SARS-CoV-2 infection, with microbial aggravation in sputum of severe asthma. *Allergy* [Internet]. 2021;76(6):1917–1921. doi:10.1111/all.14731.
105. Mu Y, Qi W, Zhang T, Zhang J, Mao S, Ishaq SL. Multi-omics analysis revealed coordinated responses of rumen microbiome and epithelium to high-grain-induced subacute Rumen acidosis in lactating dairy cows. *mSystems*. 2022;7(1):e0149021. doi:10.1128/msystems.01490-21.
  106. Sankaran K, Holmes SP. Multitable Methods for Microbiome Data Integration. *Front Genet* [Internet]. 2019;10:627. doi:10.3389/fgene.2019.00627.
  107. Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. *Mach Learn* [Internet]. 2011;83(3):331–353. doi:10.1007/s10994-010-5222-7.
  108. Park CH, Hong C, Lee A-R, Sung J, Hwang TH. Multi-omics reveals microbiome, host gene expression, and immune landscape in gastric carcinogenesis. *iSci*. 2022;25(3):103956. doi:10.1016/j.isci.2022.103956.
  109. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol* [Internet]. 2009;8(1):1–27. doi:10.2202/1544-6115.1470.
  110. Galié S, Papandreou C, Arcelin P, Garcia D, Palau-Galindo A, Gutiérrez-Tordera L, À F, Bulló M. Examining the interaction of the gut microbiome with host metabolism and cardiometabolic health in metabolic syndrome. *Nutr*. 2021;13(12):4318. doi:10.3390/nu13124318.
  111. Lisboa FJG, Peres-Neto PR, Chaer GM, da C JE, Mitchell RJ, Chapman SJ, Berbara RLL, Dalby AR. Much beyond Mantel: bringing procrustes association metric to the plant and soil ecologist's toolbox. *PLoS One* [Internet]. 2014;9(6):e101238. doi:10.1371/journal.pone.0101238.
  112. Tipton L, Cuenco KT, Huang L, Greenblatt RM, Kleerup E, Sciurba F, Duncan SR, Donahoe MP, Morris A, Ghedin E. Measuring associations between the microbiota and repeated measures of continuous clinical variables using a lasso-penalized generalized linear mixed model. *BioData Min* [Internet]. 2018;11(1):12. doi:10.1186/s13040-018-0173-9.
  113. Saji N, Niida S, Murotani K, Hisada T, Tsuduki T, Sugimoto T, Kimura A, Toba K, Sakurai T. Analysis of the relationship between the gut microbiome and dementia: a cross-sectional study conducted in Japan. *Sci Rep* [Internet]. 2019;9(1):1008. doi:10.1038/s41598-018-38218-7.
  114. Manor O, Dai CL, Kornilov SA, Smith B, Price ND, Lovejoy JC, Gibbons SM, Magis AT. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat Commun* [Internet]. 2020;11(1):5206. doi:10.1038/s41467-020-18871-1.
  115. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* [Internet]. 1996;58(1):267–288. <http://www.jstor.org/stable/2346178>. doi:10.1111/j.2517-6161.1996.tb02080.x.
  116. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* [Internet]. 2000;42(1):80–86. <http://www.jstor.org/stable/1271436>. doi:10.1080/00401706.2000.10485983.
  117. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. *J R Stat Soc Series B Stat Methodol* [Internet]. 2005 [[cited 2023 Jun 29]];67(2):301–320. <https://academic.oup.com/jrssb/article/67/2/301/7109482>.
  118. Lynch J, Tang K, Priya S, Sands J, Sands M, Tang E, Mukherjee S, Knights D, Blekhan R. HOMINID: a framework for identifying associations between host genetic variation and microbiome composition. *Gigascience* [Internet]. 2017;6(12):1–7. doi:10.1093/gigascience/gix107.
  119. Ghaemi MS, DiGiulio DB, Contrepolis K, Callahan B, Ngo TTM, Lee-McMullen B, Lehallier B, Robaczewska A, Mcilwain D, Rosenberg-Hasson Y, et al. Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. *Bioinformatics* [Internet]. 2019;35(1):95–103. doi:10.1093/bioinformatics/bty537.
  120. Mills RH, Dulai PS, Vázquez-Baeza Y, Saucedo C, Daniel N, Gerner RR, Batachari LE, Malfavon M, Zhu Q, Weldon K, et al. Multi-omics analyses of the ulcerative colitis gut microbiome link bacteroides vulgatus proteases with disease severity. *Nat Microbiol* [Internet]. 2022;7(2):262–276. doi:10.1038/s41564-021-01050-3.
  121. Gomez-Llorente MA, Martínez-Cañavate A, Chueca N, de la C RM, Romero R, Anguita-Ruiz A, Aguilera CM, Gil-Campos M, Mesa MD, Khakimov B, et al. A multi-omics approach reveals new signatures in obese allergic asthmatic children. *Biomedicines Int*. 2020;8(9):359. doi:10.3390/biomedicines8090359.
  122. Zeybel M, Arif M, Li X, Altay O, Yang H, Shi M, Akyildiz M, Saglam B, Gonenli MG, Yigit B, et al. Multiomics analysis reveals the impact of microbiota on host metabolism in hepatic steatosis. *Adv Sci* [Internet]. 2022;9(11):e2104373. doi:10.1002/advs.202104373.
  123. Bodein A, Scott-Boyer M-P, Perin O, Cao K-A L, Droit A. Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic Acids Res* [Internet]. 2022;50(5):e27. doi:10.1093/nar/gkab1200.

124. Matchado MS, Lauber M, Reitmeier S, Kacprowski T, Baumbach J, Haller D, List M. Network analysis methods for studying microbial communities: a mini review. *Comput Struct Biotechnol J* [Internet]. 2021;19:2687–2698. doi:10.1016/j.csbj.2021.05.001.
125. Perez De Souza L, Alseekh S, Brotman Y, Fernie AR. Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation. *Expert Rev Proteomics* [Internet]. 2020;17(4):243–255. doi:10.1080/14789450.2020.1766975.
126. Serin EAR, Nijveen H, Hillhorst HWM, Ligterink W. Learning from Co-expression networks: possibilities and challenges. *Front Plant Sci* [Internet]. 2016;7:444. doi:10.3389/fpls.2016.00444.
127. Loftus M, Hassouneh S-D, Yooseph S. Bacterial associations in the healthy human gut microbiome across populations. *Sci Rep* [Internet]. 2021;11(1):2828. doi:10.1038/s41598-021-82449-0.
128. Layeghifard M, Hwang DM, Guttman DS. Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol* [Internet]. 2017;25(3):217–228. doi:10.1016/j.tim.2016.11.008.
129. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform* [Internet]. 2018;19:1370–1381. doi:10.1093/bib/bbx066.
130. Liu Z, Ma A, Mathé E, Merling M, Ma Q, Liu B. Network analyses in microbiome based on high-throughput multi-omics data. *Brief Bioinform* [Internet]. 2021;22(2):1639–1655. doi:10.1093/bib/bbaa005.
131. Cantoni C, Lin Q, Dorsett Y, Ghezzi L, Liu Z, Pan Y, Chen K, Han Y, Li Z, Xiao H, et al. 2022. Alterations of host-gut microbiome interactions in multiple sclerosis. *EBioMedicine* [Internet]. 76:103798. doi: 10.1016/j.ebiom.2021.103798.
132. Pfalzer AC, Kamanu FK, Parnell LD, Tai AK, Liu Z, Mason JB, Crott JW. Interactions between the colonic transcriptome, metabolome, and microbiome in mouse models of obesity-induced intestinal cancer. *Physiol Genomics* [Internet]. 2016;48(8):545–553. doi:10.1152/physiolgenomics.00034.2016.
133. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* [Internet]. 2008;9(1):559. doi:10.1186/1471-2105-9-559.
134. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* [Internet]. 2008;40(7):854–861. doi:10.1038/ng.167.
135. Su C, Andrew A, Karagas MR, Borsuk ME. Using Bayesian networks to discover relations between genes, environment, and disease. *BioData Min* [Internet]. 2013;6(1):6. doi:10.1186/1756-0381-6-6.
136. Chella Krishnan K, Kurt Z, Barrere-Cain R, Sabir S, Das A, Floyd R, Vergnes L, Zhao Y, Che N, Charugundla S, et al. Integration of multi-omics data from mouse diversity panel highlights mitochondrial dysfunction in Non-alcoholic Fatty liver disease. *Cell Syst* [Internet]. 2018;6(1):103–15.e7. doi:10.1016/j.cels.2017.12.006.
137. Shi H, Shi Q, Grodner B, Lenz JS, Zipfel WR, Brito IL, De Vlaminck I. Highly multiplexed spatial mapping of microbial communities. *Nature* [Internet]. 2020;588(7839):676–681. doi:10.1038/s41586-020-2983-4.
138. Garg N, Wang M, Hyde E, da Silva RR, Melnik AV, Protsyuk I, Bouslimani A, Lim YW, Wong R, Humphrey G, et al. Three-dimensional microbiome and metabolome cartography of a diseased human lung. *Cell Host Microbe* [Internet]. 2017;22(5):705–16.e4. doi:10.1016/j.chom.2017.10.001.
139. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* [Internet]. 2014;505(7484):559–563. doi:10.1038/nature12820.
140. Mars RAT, Yang Y, Ward T, Houtti M, Priya S, Lekatz HR, Tang X, Sun Z, Kalari KR, Korem T, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell* [Internet]. 2020;182(6):1460–73.e17. doi:10.1016/j.cell.2020.08.007.
141. Mihindukulasuriya KA, Mars RAT, Johnson AJ, Ward T, Priya S, Lekatz HR, Kalari KR, Droit L, Zheng T, Blekhan R, et al. Multi-Omics Analyses Show disease, diet, and transcriptome interactions with the virome. *Gastroenterology* [Internet]. 2021;161(4):1194–207.e8. doi:10.1053/j.gastro.2021.06.077.
142. Ruiz-Perez D, Lugo-Martinez J, Bourguignon N, Mathee K, Lerner B, Bar-Joseph Z, Narasimhan G, Korem T. Dynamic bayesian networks for integrating multi-omics time series microbiome data. *mSystems*. 2021;6(2). doi: 10.1128/mSystems.01105-20.
143. Brumfield KD, Leddy M, Usmani M, Cotruvo JA, Tien C-T, Dorsey S, Graubics K, Fanelli B, Zhou I, Registe N, et al. Microbiome Analysis for Wastewater Surveillance during COVID-19. *MBio* [Internet]. 2022;13(4):e0059122. doi:10.1128/mbio.00591-22.
144. Herold M, Martínez Arbas S, Narayanasamy S, Sheik AR, Kleine-Borgmann LAK, Lebrun LA, Kunath BJ, Roume H, Bessarab I, Williams RBH, et al. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat Commun* [Internet]. 2020;11(1):5281. doi:10.1038/s41467-020-19006-2.
145. Muñoz-Benavent M, Hartkopf F, Van Den Bossche T, Piro VC, García-Ferris C, Latorre A, Renard BY, Muth T. gNOMO: a multi-omics pipeline for integrated host and microbiome analysis of non-model organisms. *NAR Genom Bioinform* [Internet]. 2020;2(3):lqaa058. doi:10.1093/nargab/lqaa058.

146. Blakeley-Ruiz JA, Kleiner M. Considerations for constructing a protein sequence database for metaproteomics. *Comput Struct Biotechnol J* [Internet]. 2022;20:937–952. doi:10.1016/j.csbj.2022.01.018.
147. Ankrah NYD, Bernstein DB, Biggs M, Carey M, Engevik M, García-Jiménez B, Lakshmanan M, Pacheco AR, Sulheim S, Medlock GL, et al. Enhancing microbiome research through genome-scale metabolic modeling. *mSystems* [Internet]. 2021;6(6):e0059921. doi:10.1128/mSystems.00599-21.
148. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jäger C, Baginska J, Wilmes P, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol*. 2017;35(1):81–89. doi:10.1038/nbt.3703.
149. Heinken A, Hertel J, Thiele I. Metabolic modelling reveals broad changes in gut microbial metabolism in inflammatory bowel disease patients with dysbiosis. *NPJ Syst Biol Appl* [Internet]. 2021;7(1):19. doi:10.1038/s41540-021-00178-6.
150. Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* [Internet]. 2010;26(12):i255–60. doi:10.1093/bioinformatics/btq183.
151. Kuntz TM, Gilbert JA. Introducing the microbiome into precision medicine. *Trends Pharmacol Sci* [Internet]. 2017;38(1):81–91. doi:10.1016/j.tips.2016.10.001.
152. Petrosino JF. The microbiome in precision medicine: the way forward. *Genome Med* [Internet]. 2018;10(1):12. doi:10.1186/s13073-018-0525-6.
153. Santiago-Rodriguez TM, Hollister EB. Multi ‘omic data integration: A review of concepts, considerations, and approaches. *Semin Perinatol* [Internet]. 2021;45(6):151456. doi:10.1016/j.semperi.2021.151456.
154. Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: a comprehensive review. *Front Genet* [Internet]. 2022;13:854752. doi:10.3389/fgene.2022.854752.
155. Genome Atlas Research Network C, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. *Nat Genet* [Internet]. 2013;45(10):1113–1120. doi:10.1038/ng.2764.
156. Project Consortium ENCODE, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shores N, Adrian J, Kawli T, Davis CA, Dobin A, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* [Internet]. 2020;583(7818):699–710. doi:10.1038/s41586-020-2493-4.