Cell Reports Methods

CellPress
OPEN ACCESS

Review

A practical introduction to holo-omics

Iñaki Odriozola,[1,3] Jacob A. Rasmussen,[1,3] M. Thomas P. Gilbert,[1,2] Morten T. Limborg,[1] and Antton Alberdi[1,*]
[1]Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark
[2]University Museum, NTNU, Trondheim, Norway
[3]These authors contributed equally
*Correspondence: antton.alberdi@sund.ku.dk
https://doi.org/10.1016/j.crmeth.2024.100820

SUMMARY

Holo-omics refers to the joint study of non-targeted molecular data layers from host-microbiota systems or holobionts, which is increasingly employed to disentangle the complex interactions between the elements that compose them. We navigate through the generation, analysis, and integration of omics data, focusing on the commonalities and main differences to generate and analyze the various types of omics, with a special focus on optimizing data generation and integration. We advocate for careful generation and distillation of data, followed by independent exploration and analyses of the single omic layers to obtain a better understanding of the study system, before the integration of multiple omic layers in a final model is attempted. We highlight critical decision points to achieve this aim and flag the main challenges to address complex biological questions regarding the integrative study of host-microbiota relationships.

INTRODUCTION

The realization of the importance of microorganisms for animal and plant biology,[1] along with the increased capacity to generate and process molecular data,[2] has given rise to a new holistic way to study biological systems.[3] Holo-omics refers to the technical approach to jointly (hence the prefix holo-) analyze multiple non-targeted molecular data layers (known as multi-omics) from both hosts and their associated microorganisms,[4,5] aimed at unraveling their intricate relationships.[3] The generation, analysis, and integration of holo-omic datasets requires deep knowledge of the myriad of conceptual and technical steps involved in the process.[4] Here, we provide an overview of the main steps researchers must undergo while highlighting the challenges that should be overcome to obtain meaningful results that enable them to address complex biological questions.

While acknowledging the relevance and value of other methods for studying host-microbiota interactions (e.g., 16S rRNA amplicon sequencing, spatial transcriptomics), in this review we primarily consider seven omic layers characterized by a non-targeted data generation approach without spatial resolution. These methods require specific data generation and analysis strategies before integrating them into multi-omic statistical models (Figure 1). Four of them are based on nucleic acid sequencing, namely host genomics (HG), host transcriptomics (HT), microbial metagenomics (MG), and microbial metatranscriptomics (MT). The three remaining layers are based on molecular spectroscopy, namely host proteomics (HP), microbial metaproteomics (MP), and (meta)metabolomics (ME), which is not split between hosts and microorganisms for the difficulties in assigning an origin to a given metabolite. Each omic layer contains fundamentally different information, at different levels of biological organization. A host contains a single genome (HG) that encodes thousands of genes (HT) which are expressed to produce an even larger number of proteins (HP). A host might have multiple (often hundreds or thousands) associated bacterial species (n), each with a distinct genome (MG_n) containing a few thousand genes that can be expressed (MT_n), thus potentially encoding microbial proteins (MP_n). The biological activity enabled by those proteins builds the metabolomic (ME) landscape that not only shapes host phenotype but also conditions the environment in which host-associated microbes live (e.g., the gut).

HG—which in this manuscript is solely discussed in terms of nucleotide sequences without considering other levels such as epigenetic and DNA folding that may also shape the hologenome[3,6,7]—can inform about population structure and the genetic features of individuals but is invariable to treatments or environmental disturbances. In contrast, MG contains the genetic information of microbial communities that are likely to change over very short timescales and reflects the immediate responses of hosts to treatments or disturbances. Although HG and MG inform about the potential of hosts and microorganisms to perform biological functions, HT and MT provide snapshots of the actual functional activity of the system, which can be validated using HP, MP, and ME that result from the activity of the expressed genes. Acknowledging the distinct biological characteristics of the omic layers is therefore essential to design experiments and analytical pipelines for better solving the complex puzzle of host-microbiota relationships.

In light of this, we navigate from sample acquisition to data interpretation through six sections: (1) sample collection and preservation, (2) laboratory sample processing, (3) bioinformatic analysis of raw data, (4) data filtering, imputation, and distillation, (5) initial quantitative exploration of omic layers, and (6) multi-omic data integration. Each of these sections deals with key methodological aspects, focusing on the commonalities and

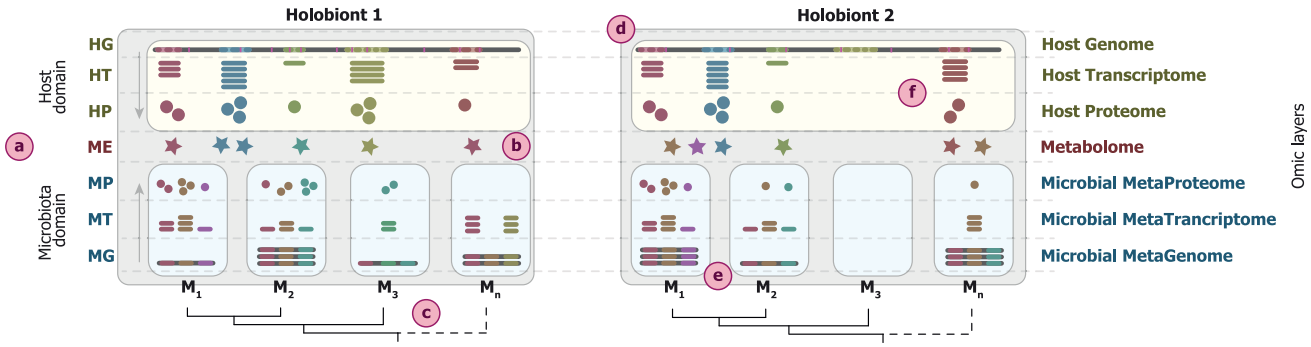

Figure 1. Schematic representation of the host-microbiota multi-omic landscape

(A) Each of the two holobionts is represented by multiple omic layers belonging to the host and microbiota domains. The metabolomic layer is left in between as it represents the molecular link between host and microbiome metabolisms and because it is often impossible to trace back the domain of origin.

(B) The host domain is depicted with a single box per holobiont representing the host organisms, whereas the microbiota domain is represented by multiple boxes, each representing a different microorganism. Note that all microorganisms present in the entire study system have been represented in both holobionts, mirroring the type of data (which includes zeros) that are used for statistical analyses.

(C) These microorganisms might have different levels of phylogenetic or functional similarities, which can be accounted for in the statistical analyses.

(D and E) Genomes are shown as dark lines containing multiple genes represented in different colors. Note that although the host genome (D) is represented with a single line, microbial genomes (E) display different numbers of genomes, representing different microbial abundances.

(F) Expressed genes and translated proteins are shown as different amounts of lines and circles that represent their quantitative nature.

main differences to generate and analyze the various types of omics, with a special focus on optimizing data generation and integration. We stress the importance of considering each step before collecting and processing data, to ensure every scientific question is aligned with the appropriate study design and the technical tools needed to address your question at hand.[3]

SAMPLE COLLECTION AND PRESERVATION

Decision-making starts with the choice of sample collection and preservation procedures, a non-trivial election that can affect the quality and accuracy of any omic datasets to the risk of masking the biological signal of interest.[8,9] Though the current "gold standard" of sample preservation is to snap-freeze and store samples at −80°C, this is not always feasible.[10] Hence, multiple different preservatives have been developed, allowing samples to be stored at ambient temperature for extended periods. Studies have shown that preservatives can introduce considerable biases when generating different omic layers,[11,12] thus keeping consistency among preservatives is critical. In addition, the biological and chemical properties of the molecules (e.g., DNA, RNA, proteins, metabolites) used to generate omic data must also be acknowledged. HG is the less sensitive omic layer because host DNA is usually abundant and stable. In contrast, MG should be treated more cautiously because microbial communities can change (e.g., the abundance of saprophytic bacteria tends to increase after sampling) unless biochemical reactions are blocked. HT and MT require even more rapid preservation if representative gene expression patterns are to be captured, because RNA typically decays faster than DNA.[13] In addition, one must bear in mind that gene expression can vary considerably across host tissues,[14] whereas microbial physiology can vary rapidly across time and space.[15] Host gene expression can vary drastically between, for instance, intestinal tissue, liver, and brain, whereas microbial gene expression can change between the mucosal layer, digesta, and feces, for example. Ultimately, selecting appropriate sample types for answering the biological question of interest in advance is essential to capture the desired signal. Lastly, metabolites have varying chemical properties, spanning large stable steroids to extremely volatile short-chain fatty acids. The choice of appropriate preservatives for the generation of multiple omic layers is therefore necessary, which entails deciding whether the omic data will be generated from a single (requires appropriate preservative for all omics) or multiple biological samples (each sample type can have a dedicated preservative). Lastly, due to the varying physicochemical properties of samples, collection and storage methods validated for one sample type cannot be assumed to be optimal for other types. Hence, preliminary optimization tests are always advisable,[16] and methodological consistency is a requirement to generate reliable omic data.

LABORATORY SAMPLE PROCESSING

Nucleic acid sequencing-based approaches

Laboratory processing of samples for DNA and RNA sequencing-based data generation (HG, HT, MG, and MT) includes extraction and purification of the nucleic acids of interest, followed by sequencing library preparation. The specific procedures to conduct these two steps can vary considerably depending on the sample type,[17] sample age,[18] and sequencing technology[19] employed. HG is the least sensitive procedure because host genomic information is largely consistent across cells and tissues.[20] By contrast, DNA extraction for MG must effectively lyse cell walls of different toughness to avoid bias toward easy-to-lyse microbes in the sequencing results.[21] RNA extraction procedures used for HT and MT are particularly sensitive, because they can introduce RNA integrity and length biases that can distort the results.[22] In addition, when dealing with low-biomass samples, such as skin swabs, one needs to

maximize the recovery of nucleic acids to capture the biological signal of interest, while minimizing the effect of contamination through the use of more strict procedures (e.g., physical separation of steps, increased cleaning effort) and control samples.

Although identical library preparation procedures can be used for HG and MG data generation,[23] HT and MT require different strategies for avoiding the domination of ribosomal RNA (rRNA) over messenger RNA (mRNA) in the resulting data. In HT and MT it is often observed that over 90% of sequences belong to rRNA transcripts unless depletion strategies are employed. The main strategy to enrich mRNA against rRNA in HT is to rely on poly(A) tails of mature transcripts of eukaryotes during the reverse transcription step to convert RNA into complementary DNA (cDNA). However, as only a small fraction of RNA harbors short oligo(A) tails in prokaryotes,[24,25] direct depletion strategies are required for MT. Historically, such rRNA depletion has been based on the hybridization of rRNA molecules with complementary oligos before library preparation.[26] Alternatively, recently developed methods based on Cas9 cleavage of repetitive molecules can be implemented after library preparation, which enables pooling multiple indexed libraries in a single reaction, thus reducing the cost and time of data generation.[27]

Molecular spectroscopy-based approaches

Tissue samples intended for nuclear magnetic resonance (NMR) spectroscopy require minimal processing, but the sensitivity and scope of protein and metabolite detection is lower compared to mass spectrometry (MS).[28] Although MS usually provides higher sensitivity over more types of molecules, it requires more complex sample preparation. Especially for highly volatile subjects, such as small metabolites, cold processing should be applied to minimize composition biases. Cell lysis by sonication or homogenization of tissue by freezing, grinding, or bead-beating is often needed to ensure efficient precipitation. Filtering of tissue samples to remove large debris, and purification of lysate for either HP, MP, or ME can be necessary to ensure high-quality data for further processing. Specifically, for ME, a solvent for the precipitation of the metabolites is needed. The solvent can range in polarity according to the desired detection spectrum of metabolites and will therefore direct the target component of the ME.[29] For HP and MP, two main approaches are currently used: one based on acetone/trichloroacetic acid precipitation and one based on phenol extraction.[30] Subsequently, for HP, MP, and ME, the removal of highly abundant metabolites or proteins can increase the detection of rare abundant proteins[31,32] and metabolites.[33,34] Because these omic layers can be quite volatile, including pooled quality controls enables the detection of all metabolites and correction of stochastic drift during the data acquisition.[35]

BIOINFORMATIC ANALYSIS OF RAW DATA

Nucleic acid data

Sequencing-based omic technologies (HG, HT, MG, and MT) produce raw data in nucleic acid sequence format with associated quality information. These quality scores are used to filter low-quality reads before downstream processing. A subsequent key decision is whether to rely on a reference-based approach where the raw reads are mapped against known reference host genomes and/or microbial metagenomes, or whether a *de novo* assembly from the generated data is necessary (Figure 2).

Reference-based approaches, often termed (meta)genome resequencing, are restricted to mapping sequencing reads to existing references. For HG, sequencing reads are typically mapped to the host reference genome using read aligners such as bwa[36] or bowtie2.[37] This process facilitates variant calling and the generation of individual genotype profiles. Though copy number variation (CNV) can be deduced from these data, analyzing structural variants typically necessitates alternative methodologies.[38,39] For MG, sequencing reads are aligned or contrasted against taxonomically or functionally annotated microbial genome databases often using Kmer-based strategies (e.g., Kraken[40]). These procedures are usually conducted using short-read sequencing data of 100–300-nucleotide-long sequences. Conversely, long-read sequencing platforms such as the PacBio or Oxford Nanopore Technologies produce thousands of nucleotides-long reads, which are particularly relevant for *de novo* genome assemblies.[19] The assembly-based approach does not require any reference data, yet it is more complex than the reference-based one. Generating high-quality host genomes requires an ensemble assembly approach that combines multiple sequencing strategies to resolve the complex structures of most eukaryotic genomes.[41] Due to this complexity, it is common practice to rely on reference host genomes, while *de novo* assembling bacterial ones. The most widely employed approach to reconstruct bacterial genomes from metagenomic mixtures is to perform a metagenomic assembly followed by contig binning.[17] Though this approach is typically executed using short-read sequences, achieving complete bacterial circular genomes from metagenomic mixtures often requires the use of long-read sequencing data.[42]

Genome assembly quality of both eukaryotic and prokaryotic organisms is assessed following similar criteria, including sequence contiguity metrics and the presence of single-copy core genes (e.g., BUSCO,[43] CheckM2[44]), but often using different software and reference information. Genome annotation is subsequently performed to identify the functional attributes of genes and other genomic elements.[45] The process yields a detailed map of the location of protein-coding genes, repetitive sequences, and other genetic elements, which are essential to obtain functional insights into host-microbiota interactions in downstream analyses.

The dichotomy between reference- or assembly-based strategies also applies to HT and MT (Figure 3). If relevant HG and MG data are already available, then (meta)transcriptomic data can be simply mapped against (meta)genomic sequences such as a reference host genome, a custom microbial genome catalog, or a public microbial genome database (e.g., Chocophlan[46]) to quantify gene expression. This step must be performed to acknowledge the structural differences between eukaryotic (contain introns) and prokaryotic (do not contain introns) genes because splice-aware read mappers are required for the former. However, when a suitable quality reference genome is unavailable, researchers may opt to assemble transcripts and assign functions based on reference gene libraries, which adhere to the same principles as HG and MG, followed by mapping of

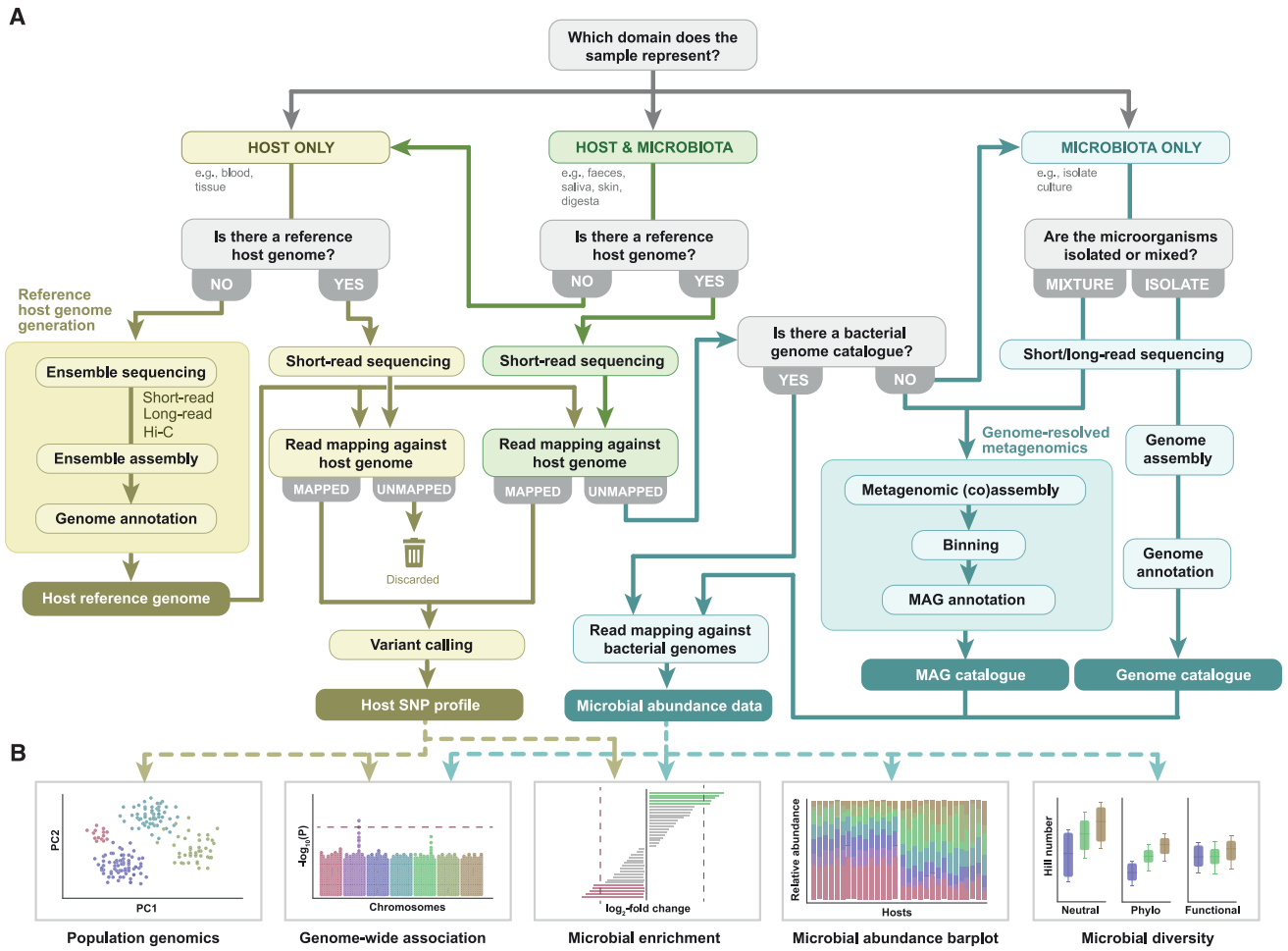

Figure 2. Overview of data generation and analysis pathways for host genomics and microbial metagenomics

(A) The datatype to be generated and the procedures to be followed depend on prior data availability and the sample type employed. Procedures dealing with host-only, microbe-only, or combined data are shown in different colors.

(B) Examples of statistical analyses that can be carried out using host-only, microbe-only, and combined information, after filtration, distillation, and transformation of the data.

reads to the assembled transcripts to quantify their expression profiles.

Molecular spectroscopy data

Compared to nucleic acids, proteins and metabolites are more complex molecules that are most often characterized by a large number of spectra generated by spectrometers that require processing before meaningful biological inferences can be made. The most widely used form of data collection in quantitative proteomics and metabolomics when using tandem mass spectrometry (MS/MS), is quantified based on the first spectrometer (MS^1-level) and identified at the second spectrometer (MS^2-level). MS^2-spectra, in turn, originate from the fragmentation of precursors observed in MS^1-spectra and are usually selected based on peak height and charge state. Quantitative MS is normally based on the area under the curve (AUC) or peak height calculation for each feature that elutes from the chromatograph column at an expected retention time. These metrics can then be employed to quantify the corresponding features in HP, MP, and ME using the detection from MS^1.[47–50] Subsequently, feature identification is achieved through MS^2. Here, search algorithms aim to explain a recorded MS^2 spectrum by a feature spectrum from a predefined database, returning a list of features that fit the experimental data with a certain confidence score.

HP and MP databases are normally protein databases translated from genomic data,[51] although other strategies such as spectral libraries[52] or mRNA databases[53] have been successfully applied. Assembling the identified peptides into proteins can be challenging, particularly when dealing with redundant peptides or spliced proteins.[54,55] On the other hand, recent computational advances for the prediction of protein structures will continue to increase reference databases for proteomics.[56–58] Because ME cannot rely on databases translated from genomic data, different strategies are needed. The most common practice for the identification of specific metabolites has been to use custom libraries of known metabolites.

CellPress
OPEN ACCESS

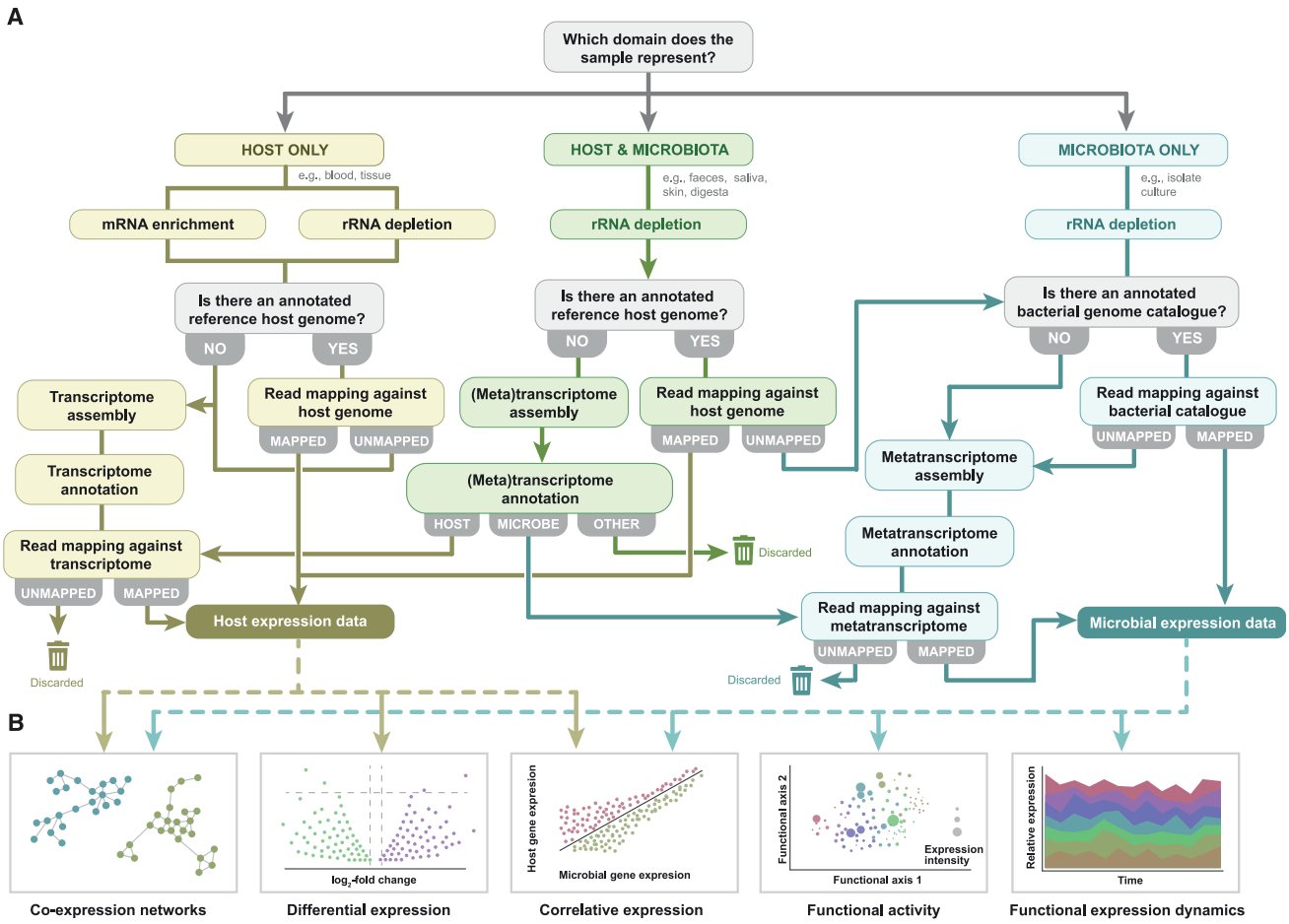

Figure 3. Overview of data generation and analysis pathways for host transcriptomics and microbial metatranscriptomics

(A) The datatype to be generated and the procedures to be followed depend on prior data availability and sample type employed. Procedures dealing with host-only, microbe-only, or combined data are shown in different colors.

(B) Examples of statistical analyses that can be carried out using host-only, microbe-only, and combined information, after filtration, distillation, and transformation of the data.

However, due to the high costs of preparing custom libraries, community-driven strategies such as metabolic feature databases have started to emerge.[59,60] Recent advances toward feature-based molecular networks (FBMNs) have also increased the *in silico* annotation of unknown metabolic features[61–63] (Figure 4).

DATA FILTERING, IMPUTATION, AND DISTILLATION

Bioinformatic processing of raw sequencing and spectrometry data yields a massive amount of information organized in structurally diverse datasets. Host genomes often contain thousands of genes, with hundreds of thousands of nucleotide variants.[23] It is also common to generate catalogs of hundreds of bacterial genomes, with millions of genes, as well as metabolomic profiles containing thousands of metabolites. One of the first considerations is to filter these data, to reduce the information to only that required for answering the biological question of interest. Following qualitative criteria, one may decide to exclude entire omic layers to address specific questions or to only consider a subset of an omic dataset, such as a specific family of genes. Following quantitative criteria, it is necessary to establish thresholds to discard data points under certain threshold values to offset the biases and limitations of data generation techniques. Insufficient sequencing effort, genome completeness, or coverage of genomes are some of the reasons researchers need to filter some of the data out, to avoid technical aspects introducing noise that potentially shadows the true biological signal in the statistical analyses.

As many statistical approaches do not deal adequately with missing data, imputation is another strategy to be considered in the initial pre-processing of the data. Imputation refers to the process of replacing missing data with the most probable substituted values based on the context. Genotype imputation is broadly employed in the context of HG,[23] functional imputation has also been recently proposed as an approach to minimize bias from low genome completeness in functional MG analyses,[64,65] and imputation strategies are also being developed for HT and ME.[66,67]

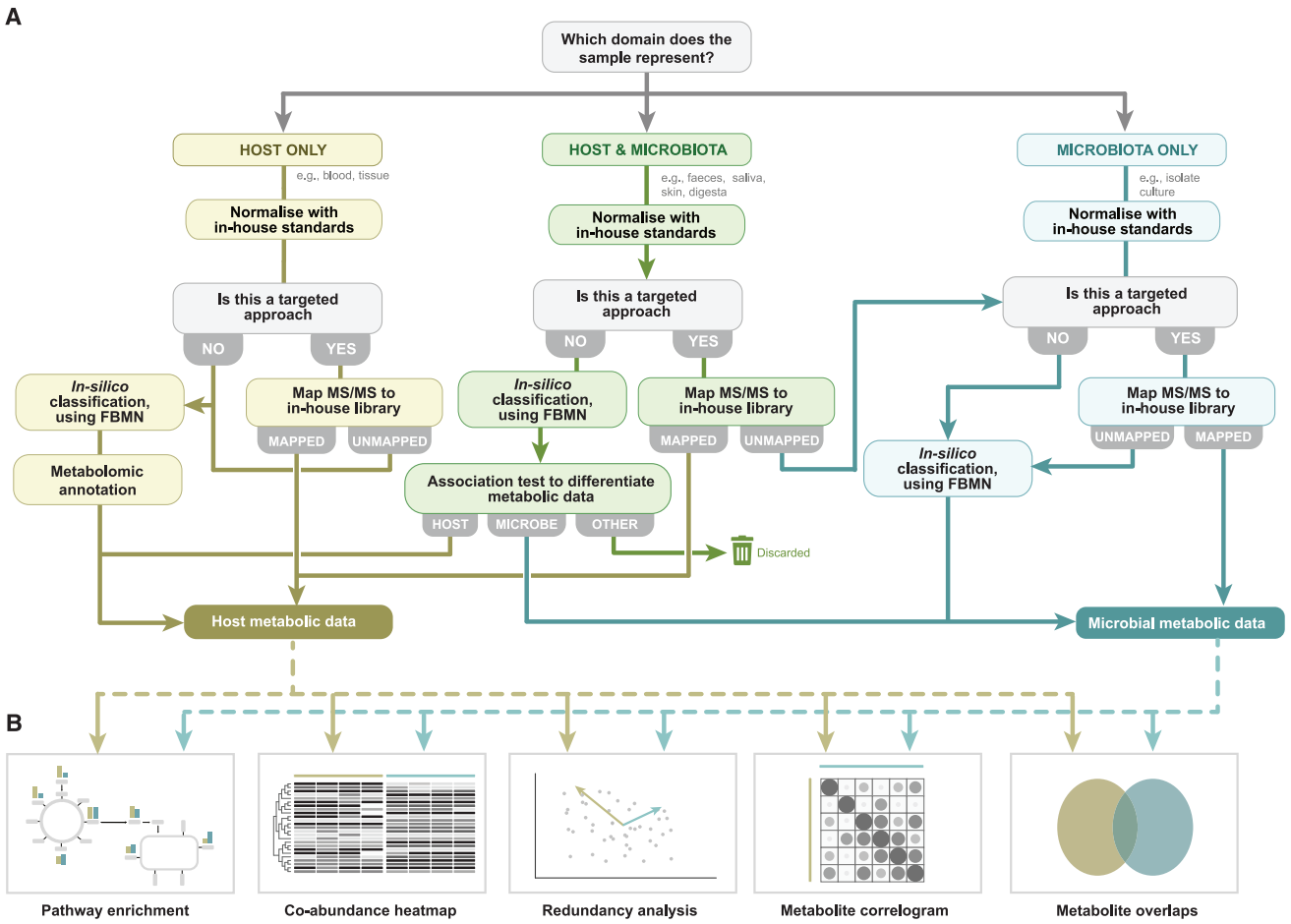

Figure 4. Overview of data generation and analysis pathways for metabolomics

(A) The datatype to be generated and the procedures to be followed depend on prior data availability and the sample type employed. Procedures dealing with host-only, microbe-only, or combined data are shown in different colors.

(B) Examples of statistical analyses that can be carried out using metabolomic data.

In some cases, integration and interpretation of these multivariable data can be intractable, which might necessitate the conversion of original datasets into a reduced set of biologically meaningful data through a process known as data distillation.[45] The main advantage of data distillation is to gain the capacity to interpret results within a given biological context. Moreover, distillation also reduces the number of features, thus decreasing computational requirements and gaining statistical power for downstream analyses. One example of biological data distillation is to convert occurrence data of dozens of genes present in a bacterial genome into a genome-inferred functional trait, which proxies the capacity of that genome to perform a given function.[68]

INITIAL QUANTITATIVE EXPLORATION OF OMIC LAYERS

The analysis of multi-omic data should begin with an independent analysis of each omic layer to learn about its structure and variability before jumping to multi-omic data integration. Despite the fundamental differences between the seven omic layers considered in this article, all share the characteristic of being multivariate, i.e., they contain multiple features (genomic variants, genes, metabolic pathways, proteins, or metabolites) collected across multiple observations. In the following, we provide a summary of some of the most relevant statistical techniques suitable for multi-omic studies, which have been covered more in depth in the literature.[69–72]

Data transformations

Multivariate datasets often consist of different data types (e.g., presence-absence of taxa, counts of genes, concentrations of metabolites across samples) that may require specific transformation before applying statistical techniques. Moreover, many of the statistical techniques applied to those datasets make assumptions such as sampled values being normally (or at least symmetrically) distributed, variables being independent of each other, or population variances being homogeneous. Biological datasets rarely meet these assumptions, but original values can be transformed so that the modified values conform

better to those assumptions.[73] Typical transformations to change variable distributions include log transformations (with a small value added to deal with zeros), root- ransformations (when distributions are right-skewed), square ransformations (when distributions are left-skewed), or arcsine transformation (applied to fractional or proportional data). Additionally, despite the aforementioned possibilities for data imputation, biological datasets often contain missing observations (e.g., genes, genotypic variants, or genomes) recorded as zeros. Thus, many common euclidean-based statistical tools such as principal-component analysis (PCA), redundancy analysis (RDA), or K-means clustering are not suitable for such datasets. To apply these methods to zero-inflated presence-absence and abundance datasets, another set of transformations (e.g., Hellinger) is recommended.[73]

Another aspect that must be considered when working with multi-omic datasets is whether measurements represent absolute or relative values. Although HG provides qualitative information of host genomes, MG, HT, and MT yield quantitative information that depends on the amount of sequencing performed, i.e., they are compositional. Consequently, raw quantitative values of genome abundance or gene expression measured across sampling units cannot be directly compared. The most common solution is to normalize raw abundance values into relative abundance data, which allows comparing values of the features across observations. However, this normalization causes individual variables to lose their independence, which is one of the assumptions in many statistical methods.[71] Alternatively, ratio transformations such as the centered-log-ratio have been proposed as more appropriate for compositional data analysis, because they remove the effect of constant-sum constraint on the covariance and correlation matrices.[71,74,75]

Finally, standardization (i.e., scaling to mean of zero and unit variance) is often needed to avoid the excessive influence of certain features due to their larger magnitudes. For instance, in ME certain metabolites, such as ATP, may show much higher concentrations than other essential metabolites, such as signaling molecules, and potentially swamp interesting differences among less abundant, but more meaningful, metabolites.[76] Similarly, in HT and MT, transcript-length biases can also produce similar distortions.[77] However, inflating the influence of small values through standardization may also increase the influence of the measurement error, which is usually relatively larger for less abundant features.[76]

Unsupervised exploration of omic layers

Unsupervised methods include exploratory techniques, such as cluster analysis and ordination-based visualization methods, which reveal the structure and main patterns of the omic datasets without prior information about experimental design. These procedures might reveal that the observations are structured into biologically meaningful groups or that variables can be reduced to fewer dimensions. Researchers can then opt for using the outputs of these analyses, rather than the original datasets, for the multi-omic data integration to reduce the complexity and size of the data. As most clustering and ordination techniques are computed from association matrices, appropriate pre-transformation of the data and choice of association coefficient are essential, since these influence the final outcome of the analyses.[69]

Cluster analysis

Clustering procedures group features or observations into homogeneous sets by minimizing within-group and maximizing among-group distances.[71] This is commonly achieved either based on hierarchical or disjoint methods.[71] Hierarchical clustering produces a stratified organization of features or observations where relatively similar objects are grouped.[69,71] A useful exploratory analysis to reveal general patterns in an omic layer can be obtained by simultaneous application of hierarchical clustering to the rows and columns of the data matrix, and visualizing the results in a heatmap. Disjoint clustering techniques aim at separating the objects into individual, usually mutually exclusive, and in most cases, unconnected clusters.[71] K-means clustering is one of the most broadly employed algorithms, although many other methods are also available.

Dimension reduction and ordination

Ordination is a complementary method to data clustering, which enables displaying differences among samples graphically through reducing the dimensions of the original dataset so that similar objects are near and dissimilar objects are farther from each other. PCA, principal-coordinate analysis (PCoA), and non-metric multidimensional scaling (NMDS) are three of the most popular methods used for ordination.[78] More recent machine learning (ML) algorithms for dimensionality reduction may achieve better representations of complex non-linear relationships in omic datasets if their parameters are properly tuned.[79] These include t-distributed stochastic neighbor embedding (t-SNE)[80] and uniform manifold approximation and projection (UMAP),[81] among others.

Supervised analysis of omic layers

The supervised analyses of omic layers, in contrast to unsupervised ones, incorporate information on experimental design and can be divided into two types of problems: regression and classification. A regression problem is when the output of the model is a numeric variable or a matrix, such as the phenotypic characteristics of the host or the omic datasets themselves. These methods aim at testing and estimating the effects of the experimental factors (e.g., dietary treatment, drug administration) or variables of interest (e.g., age of the experimental subjects, geographic location of studied populations) on different omic layers, or at associating the omic layers with host phenotypic features. A classification problem is when the output of the model is categorical. In the context of multi-omic studies, classification methods aim at classifying observations into their experimental groups (e.g., health status, dietary treatment) based on their features on different omic layers.

Regression methods

Independently testing the effects of the experimental factors of interest on different omic layers can be very informative for getting an overall picture of how the host and the microbiome are responding to the environment and/or the experimental treatment. Popular methods to test the effects of explanatory variables on multivariate response tables are PERMANOVA,[82] ANOSIM[83] and, although more restricted in the association measures assumed among objects, RDA and canonical correspondence

analysis (CCA). Generalized linear modeling (GLM) is probably the most popular framework for regression analysis with omic layers. GLMs are an extension of standard linear regression and analysis of variance (ANOVA), with the advantage that not only continuous but also discrete, binary, or proportional outcomes can be used, including zero-inflated data.[84] Moreover, the extensions to generalized linear mixed modeling (GLMM) allow analyses of complex experimental designs with hierarchical, spatial, and temporal structures[85] and extensions to generalized additive mixed modeling (GAMM) relax the assumption of linearity and allow fitting complex non-linear relationships between outcomes and explanatory variables.[86] Such approaches have been extensively employed to find associations between host genotypes and microbial metagenotypes, and host phenotypes such as production parameters and disease states. Recent developments in joint species distribution modeling have extended the flexibility of the GLMM framework to accommodate multivariate data tables as response variables.[87] Although other ML algorithms can also be used for regression problems,[88–91] GLM-based methods are used more often in regression. This is because of the convenience of interpreting an output as a linear combination of inputs, and also because often GLM-based methods obtain comparable performance to more complex algorithms.[79]

Classification methods

Traditionally, classification problems have been solved using linear-based methods such as linear discriminant analysis or logistic regression (if the response variable is binary, e.g., healthy vs. sick). However, it is in classification problems where ML algorithms such as random forests (RFs),[92] gradient boosting machines (GBMs),[93] or support vector machines (SVMs)[94] have proven most useful thanks to their capacity to handle a high number of features concerning observations.[95] In recent years, deep learning (DL) has emerged as a promising class of techniques that has beaten the performance of more classic algorithms in multiple classification problems, from image and speech recognition to biological function prediction.[96] DL is especially well suited to analyze large-scale datasets with high dimensionality, such as the ones generated in multi-omic studies. However, DL algorithms are computationally expensive and are often referred to as ''black box'' algorithms, in the sense that it is challenging to identify the features that a neural network deems most significant for classification.[97]

MULTI-OMIC DATA INTEGRATION

When it comes to multi-omic data integration, researchers often need to prioritize between predicting a given outcome (e.g., phenotype, disease state, microbial community) based on the multi-omic data (predictive modeling) or using multi-omic data for understanding the biological processes leading to that outcome (explanatory modeling or causal inference). A good predictive model includes a set of variables, the observation of which is systematically associated with changes in an outcome, thus maximizing predictive capacity. By contrast, a good causal model includes a set of variables, the intervention of which would provoke a change in an outcome.[98] Both concepts are related in that a good causal model should have predictive capacity; however, the best predictive model need not have a causal meaning.[99]

Acknowledging these priorities is important for choosing the best strategy for multi-omic data integration, which can be broadly categorized into two types: multi-staged analysis and meta-dimensional analysis.[100,101] In the multi-staged approximation, data analysis is divided into multiple steps, relating two omic layers at a time, and the final step links the relevant omic layers with the outcome of interest.[100] This approach leverages the central dogma of molecular biology to assume that the variation in omic datasets is hierarchical, such that variation in DNA leads to variation in RNA and so on. By contrast, in meta-dimensional analysis, all omic datasets are analyzed in a single analysis, spanning the whole landscape of features simultaneously and enabling to account for inter-omic interactions.[100]

Until recently, the multi-staged approach dominated the field of integrated analysis of biological data,[101] mostly relying on traditional statistical tools and hypothesis testing approximations mentioned in the previous section. Multi-staged analysis has the advantage of relating multi-omic datasets in an organized, stepwise manner, which enables building knowledge for later use to test causally oriented hypotheses. In addition, multi-staged approaches are better suited to accommodate the biological asymmetries between multi-omic datasets. For instance, although both HG and MG data comprise DNA sequences that could be treated similarly, the process that generated their compositions profoundly differs. While HG belongs to a single individual (i.e., the host), the MG belongs to a microbial community of multiple individuals. Species communities are assembled following the basic processes of selection, drift, dispersal, and speciation,[102] the combination of which determines the community composition, functional profile, and biodiversity of the microbiome. Thus, inferences made from MG may benefit from specific statistical frameworks such as joint species distribution modeling,[87,103] whereas this might not be the case for HG data. As a downside, such a stepwise approach assumes that omic layers linearly influence each other to determine a complex trait of interest, which implies that complex inter-omic interactions are likely to be overlooked.[100]

Advances in the field of artificial intelligence (AI) and ML have shifted the balance toward meta-dimensional approaches where the whole complexity of multi-omic datasets are pooled into a single analysis.[97] Meta-dimensional approaches can be classified into concatenation-based, transformation-based, and model-based integration methods.[97,100] Concatenation-based integration combines multiple omic datasets, raw or pre-processed, into a single large matrix; then, the concatenated table can be used for unsupervised learning with methods such as multi-omics factor analysis (MOFA),[104] or for supervised learning with any of the above-mentioned ML algorithms. In transformation-based integration, omic datasets are first transformed into an intermediate representation, typically a graph or a kernel matrix, and they are then merged before building the final model. Graph-based analyses have the advantage of easier interpretability and lower computational requirements whereas, overall, kernel-based methods provide higher predictive performance.[105] Model-based integration builds intermediate models from each omic layer and then builds a final model combining

all intermediate models. An advantage of the model-based approach is that it allows the merging of multiple omic types that have been collected in different sets of sampling units if the outcome of interest is the same across datasets (e.g., specific disease).[100] Multi-omic data integration efforts such as analysis tool for heritable and environmental network associations (ATHENA)[106] or multi-omics supervised autoencoder (MOSAE)[107] use model-based integration for disease prediction by combining a variety of modeling frameworks and algorithms. Unquestionably, meta-dimensional integration of multi-omic data has improved predictive accuracy in many applications, including disease diagnostics and prognosis or biomarker discovery.[108,109] However, these applications usually respond to rather "simple" problems, such as classifying individuals into health statuses.[97] Such simple settings are the exception rather than the rule in studies on fundamental biology, which often deal with intricate hierarchical, spatial, temporal, or phylogenetic structures derived from complex experimental designs and the heterogeneity of nature. Although approaches are being developed to account for those biases in ML algorithms,[110,111] currently generalized linear mixed models and Bayesian hierarchical models represent the most established and accessible statistical frameworks to account for them by means of random effects.

Lastly, one must bear in mind that predictive modeling is not suited for causal inference, because non-causal associations may increase predictive accuracy, but undermine our understanding of the system under study.[112,113] Unlike the multi-staged approximation, meta-dimensional analysis favors not using previous knowledge to independently reduce omic datasets, arguing that domain-knowledge-oriented procedures may introduce bias by ignoring previously undiscovered biology.[100] However, blindly pooling hundreds of variables into a model may introduce biases as well, such as masking true causal associations through overcontrol bias or generating spurious (i.e., non-causal) correlations through collider bias.[114] Hence, while maximizing predictive accuracy is highly valuable for many applications of multi-omic data, we argue it should not be the final goal when trying to understand complex host-microbe and microbe-microbe interactions in basic holo-omic research. In those situations, the application of meta-dimensional approaches should ideally be an intermediate step to generate hypotheses on causal relationships that later will need to be tested using randomized experiments or causally oriented analyses, which have recently started to flourish in the context of multi-omic data integration.[115,116]

CONCLUSIONS AND FUTURE PERSPECTIVES

The number of elements involved, the asymmetry between host and microbial organisms, the dynamism of the interactions, and the multi-scale nature of the effects render host-microbiota systems, or holobionts, as one of the most challenging biological entities to be studied. Holo-omics tackles that complexity by leveraging the full potential of laboratory, bioinformatic, and statistical methodologies that continue advancing at a vertiginous pace. Although access to great data and technology comes with great power, with great power comes great responsibility. In certain scenarios, employing simpler targeted methodologies, such as characterizing taxonomic compositions of microbial communities through 16S rRNA sequencing or quantifying specific metabolites such as short-chain fatty acids, could prove more cost effective than adopting non-targeted approaches. However, for more complex inquiries, higher resolution strategies such as epigenomics, single-cell genomics, Hi-C genomics, spatial transcriptomics, or spatial metabolomics might be necessary to effectively address research questions demanding a complete representation of the environment. In any case, generating millions of data points does not entail that all of them need to be used without any filtering or distillation. Instead, we argue that the workflow delineated in this article, with careful generation and distillation of data followed by independent exploration and analyses of the single omic layers, will aid researchers in having a better understanding of the study system before the integration of multiple omic layers in a final model is attempted. We, therefore, advocate for an integral approximation in which study design (acknowledging the experimental setup that best allows addressing the problem of interest), data generation (acknowledging the pitfalls and biases of employed techniques), and data analysis (acknowledging the properties of the biological elements under study) are considered as a whole. Only then will we be able to maximize the power of holo-omics and address some of the most challenging questions in modern biology.

ACKNOWLEDGMENTS

This work was supported by the Danish National Research Foundation under grant DNRF143 "A Center for Evolutionary Hologenomics", and the European Union's Horizon Research and Innovation Programme under grant no. 817729.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. McFall-Ngai, M., Hadfield, M.G., Bosch, T.C.G., Carey, H.V., Domazet-Lošo, T., Douglas, A.E., Dubilier, N., Eberl, G., Fukami, T., Gilbert, S.F., et al. (2013). Animals in a bacterial world, a new imperative for the life sciences. Proc. Natl. Acad. Sci. USA *110*, 3229–3236.
2. Giani, A.M., Gallo, G.R., Gianfranceschi, L., and Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput. Struct. Biotechnol. J. *18*, 9–19.
3. Alberdi, A., Andersen, S.B., Limborg, M.T., Dunn, R.R., and Gilbert, M.T.P. (2022). Disentangling host–microbiota complexity through hologenomics. Nat. Rev. Genet. *23*, 281–297.
4. Nyholm, L., Koziol, A., Marcos, S., Botnen, A.B., Aizpurua, O., Gopalakrishnan, S., Limborg, M.T., Gilbert, M.T.P., and Alberdi, A. (2020). Holo-Omics: Integrated Host-Microbiota Multi-omics for Basic and Applied Biological Research. iScience *23*, 101414.
5. Xu, L., Pierroz, G., Wipf, H.M.-L., Gao, C., Taylor, J.W., Lemaux, P.G., and Coleman-Derr, D. (2021). Holo-omics for deciphering plant-microbiome interactions. Microbiome *9*, 69.
6. Hansen, S.B., Bozzi, D., Mak, S.S.T., Clausen, C.G., Nielsen, T.K., Kodama, M., Hansen, L.H., Gilbert, M.T.P., and Limborg, M.T. (2023). Intestinal epigenotype of Atlantic salmon (Salmo salar) associates with tenacibaculosis and gut microbiota composition. Genomics *115*, 110629.

7. Zhang, H., Kalla, R., Chen, J., Zhao, J., Zhou, X., Adams, A., Noble, A., Ventham, N.T., Wellens, J., Ho, G.-T., et al. (2024). Altered DNA methylation within DNMT3A, AHRR, LTA/TNF loci mediates the effect of smoking on inflammatory bowel disease. Nat. Commun. *15*, 595.

8. Hamady, M., and Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. *19*, 1141–1152.

9. Lozupone, C.A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., Jansson, J.K., Gordon, J.I., and Knight, R. (2013). Meta-analyses of studies of the human microbiota. Genome Res. *23*, 1704–1714.

10. Song, S.J., Amir, A., Metcalf, J.L., Amato, K.R., Xu, Z.Z., Humphrey, G., and Knight, R. (2016). Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. mSystems *1*, e00021-16. https://doi.org/10.1128/mSystems.00021-16.

11. Gratton, J., Phetcharaburanin, J., Mullish, B.H., Williams, H.R.T., Thursz, M., Nicholson, J.K., Holmes, E., Marchesi, J.R., and Li, J.V. (2016). Optimized Sample Handling Strategy for Metabolic Profiling of Human Feces. Anal. Chem. *88*, 4661–4668.

12. Choo, J.M., Leong, L.E.X., and Rogers, G.B. (2015). Sample storage conditions significantly influence faecal microbiome profiles. Sci. Rep. *5*, 16350.

13. Kagzi, K., Hechler, R.M., Fussmann, G.F., and Cristescu, M.E. (2022). Environmental RNA degrades more rapidly than environmental DNA across a broad range of pH conditions. Mol. Ecol. Resour. *22*, 2640–2650.

14. Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S.E., Hamel, A.R., Viñuela, A., Roberts, A.L., et al. (2020). Cell type–specific genetic regulation of gene expression across human tissues. Science *369*, eaaz8528.

15. Donaldson, G.P., Chou, W.-C., Manson, A.L., Rogov, P., Abeel, T., Bochicchio, J., Ciulla, D., Melnikov, A., Ernst, P.B., Chu, H., et al. (2020). Spatially distinct physiology of Bacteroides fragilis within the proximal colon of gnotobiotic mice. Nat. Microbiol. *5*, 746–756.

16. Hildonen, M., Kodama, M., Puetz, L.C., Gilbert, M.T.P., and Limborg, M.T. (2019). A comparison of storage methods for gut microbiome studies in teleosts: Insights from rainbow trout (Oncorhynchus mykiss). J. Microbiol. Methods *160*, 42–48.

17. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833–844.

18. McGaughran, A. (2020). Effects of sample age on data quality from targeted sequencing of museum specimens: what are we capturing in time? BMC Genom. *21*, 188.

19. Tedersoo, L., Albertsen, M., Anslan, S., and Callahan, B. (2021). Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. Appl. Environ. Microbiol. *87*, e0062621.

20. Aizpurua, O., Dunn, R.R., Hansen, L.H., Gilbert, M.T.P., and Alberdi, A. (2023). Field and laboratory guidelines for reliable bioinformatic and statistical analysis of bacterial shotgun metagenomic data. Crit. Rev. Biotechnol., 1–19.

21. Yuan, S., Cohen, D.B., Ravel, J., Abdo, Z., and Forney, L.J. (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS One *7*, e33865.

22. Byrne, A., Cole, C., Volden, R., and Vollmers, C. (2019). Realizing the potential of full-length transcriptome sequencing. Philos. Trans. R. Soc. Lond. B Biol. Sci. *374*, 20190097.

23. Marcos, S., Parejo, M., Estonba, A., and Alberdi, A. (2022). Recovering high-quality host genomes from gut metagenomic data through genotype imputation. Adv. Genet. *3*, 2100065.

24. Dreyfus, M., and Régnier, P. (2002). The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. Cell *111*, 611–613.

25. Sarkar, N. (1997). Polyadenylation of mRNA in prokaryotes. Annu. Rev. Biochem. *66*, 173–197.

26. Huang, Y., Sheth, R.U., Kaufman, A., and Wang, H.H. (2020). Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics. Nucleic Acids Res. *48*, e20.

27. Prezza, G., Heckel, T., Dietrich, S., Homberger, C., Westermann, A.J., and Vogel, J. (2020). Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. RNA *26*, 1069–1078.

28. Emwas, A.-H.M. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. Methods Mol. Biol. *1277*, 161–193.

29. Vuckovic, D. (2012). Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. Anal. Bioanal. Chem. *403*, 1523–1548.

30. Wang, W., Tai, F., and Chen, S. (2008). Optimizing protein extraction from plant tissues for enhanced proteomics analysis. J. Sep. Sci. *31*, 2032–2039.

31. Kim, Y.J., Wang, Y., Gupta, R., Kim, S.W., Min, C.W., Kim, Y.C., Park, K.H., Agrawal, G.K., Rakwal, R., Choung, M.-G., et al. (2015). Protamine sulfate precipitation method depletes abundant plant seed-storage proteins: A case study on legume plants. Proteomics *15*, 1760–1764.

32. Gupta, R., and Kim, S.T. (2015). Depletion of RuBisCO protein using the protamine sulfate precipitation method. Methods Mol. Biol. *1295*, 225–233.

33. Rico, E., González, O., Blanco, M.E., and Alonso, R.M. (2014). Evaluation of human plasma sample preparation protocols for untargeted metabolic profiles analyzed by UHPLC-ESI-TOF-MS. Anal. Bioanal. Chem. *406*, 7641–7652.

34. Michopoulos, F., Lai, L., Gika, H., Theodoridis, G., and Wilson, I. (2009). UPLC-MS-based analysis of human plasma for metabonomics using solvent precipitation or solid phase extraction. J. Proteome Res. *8*, 2114–2121.

35. Fiehn, O., Wohlgemuth, G., Scholz, M., Kind, T., Lee, D.Y., Lu, Y., Moon, S., and Nikolau, B. (2008). Quality control for plant metabolomics: reporting MSI-compliant studies. Plant J. *53*, 691–704.

36. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

37. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

38. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

39. Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. *19*, 453–467.

40. Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

41. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746.

42. Kim, C., Pongpanich, M., and Porntaveetus, T. (2024). Unraveling metagenomics through long-read sequencing: a comprehensive review. J. Transl. Med. *22*, 111.

43. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

44. Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat. Methods *20*, 1203–1212. https://doi.org/10.1038/s41592-023-01940-w.

45. Shaffer, M., Borton, M.A., McGivern, B.B., Zayed, A.A., La Rosa, S.L., Solden, L.M., Liu, P., Narrowe, A.B., Rodríguez-Ramos, J., Bolduc, B.,

et al. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. *48*, 8883–8900.

46. Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife *10*, e65088. https://doi.org/10.7554/eLife.65088.
47. Huan, T., and Li, L. (2015). Quantitative Metabolome Analysis Based on Chromatographic Peak Reconstruction in Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry. Anal. Chem. *87*, 7011–7016.
48. Kapoore, R.V., and Vaidyanathan, S. (2016). Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems. Philos. Trans. A Math. Phys. Eng. Sci. *374*, 20150363. https://doi.org/10.1098/rsta.2015.0363.
49. Li, Y., and Li, L. (2018). Improving accuracy of peak-pair intensity ratio measurement in differential chemical isotope labeling LC–MS for quantitative metabolomics. Int. J. Mass Spectrom. *434*, 202–208.
50. Rozanova, S., Barkovits, K., Nikolov, M., Schmidt, C., Urlaub, H., and Marcus, K. (2021). Quantitative Mass Spectrometry-Based Proteomics: An Overview. In Quantitative Methods in Proteomics, K. Marcus, M. Eisenacher, and B. Sitek, eds. (Springer US), pp. 85–116.
51. Kumar, C., and Mann, M. (2009). Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Lett. *583*, 1703–1712.
52. Lam, H. (2011). Building and Searching Tandem Mass Spectral Libraries for Peptide Identification. Mol. Cell. Proteomics *10*, R111.008565.
53. Wang, X., Slebos, R.J.C., Wang, D., Halvey, P.J., Tabb, D.L., Liebler, D.C., and Zhang, B. (2012). Protein identification using customized protein sequence databases derived from RNA-Seq data. J. Proteome Res. *11*, 1009–1017.
54. Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics *4*, 1419–1440.
55. Plubell, D.L., Käll, L., Webb-Robertson, B.-J., Bramer, L.M., Ives, A., Kelleher, N.L., Smith, L.M., Montine, T.J., Wu, C.C., and MacCoss, M.J. (2022). Putting Humpty Dumpty Back Together Again: What Does Protein Quantification Mean in Bottom-Up Proteomics? J. Proteome Res. *21*, 891–898.
56. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.
57. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444.
58. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130.
59. Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol. *34*, 828–837.
60. Haug, K., Cochrane, K., Nainala, V.C., Williams, M., Chang, J., Jayaseelan, K.V., and O'Donovan, C. (2020). MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res. *48*, D440–D444.
61. van der Hooft, J.J.J., Wandy, J., Barrett, M.P., Burgess, K.E.V., and Rogers, S. (2016). Topic modeling for untargeted substructure exploration in metabolomics. Proc. Natl. Acad. Sci. USA *113*, 13738–13743.
62. da Silva, R.R., Wang, M., Nothias, L.-F., van der Hooft, J.J.J., Caraballo-Rodríguez, A.M., Fox, E., Balunas, M.J., Klassen, J.L., Lopes, N.P., and Dorrestein, P.C. (2018). Propagating annotations of molecular networks using in silico fragmentation. PLoS Comput. Biol. *14*, e1006089.
63. Ernst, M., Kang, K.B., Caraballo-Rodríguez, A.M., Nothias, L.-F., Wandy, J., Chen, C., Wang, M., Rogers, S., Medema, M.H., Dorrestein, P.C., et al. (2019). MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. Metabolites *9*, 144. https://doi.org/10.3390/metabo9070144.
64. Geller-McGrath, D., Konwar, K.M., Edgcomb, V.P., Pachiadaki, M., Roddy, J.W., Wheeler, T.J., and McDermott, J.E. (2023). MetaPathPredict: A machine learning-based tool for predicting metabolic modules in incomplete bacterial genomes. Preprint at bioRxiv. https://doi.org/10.1101/2022.12.21.521254.
65. Eisenhofer, R., Odriozola, I., and Alberdi, A. (2023). Impact of microbial genome completeness on metagenomic functional inference. ISME Commun. *3*, 12.
66. Patruno, L., Maspero, D., Craighero, F., Angaroni, F., Antoniotti, M., and Graudenzi, A. (2021). A review of computational strategies for denoising and imputation of single-cell transcriptomic data. Brief. Bioinform. *22*, bbaa222. https://doi.org/10.1093/bib/bbaa222.
67. Shahjaman, M., Rahman, M.R., Islam, T., Auwul, M.R., Moni, M.A., and Mollah, M.N.H. (2021). rMisbeta: A robust missing value imputation approach in transcriptomics and metabolomics data. Comput. Biol. Med. *138*, 104911.
68. Koziol, A., Odriozola, I., Leonard, A., Eisenhofer, R., San José, C., Aizpurua, O., and Alberdi, A. (2023). Mammals show distinct functional gut microbiome dynamics to identical series of environmental stressors. mBio *14*, e0160623. https://doi.org/10.1128/mbio.01606-23.
69. Legendre, P., and Legendre, L. (2012). Numerical Ecology, 3rd English Edition 3rd ed. (Elsevier science).
70. Borcard, D., Gillet, F., and Legendre, P. (2011). In Numerical Ecology with, R.R. Gentleman, K. Hornik, and G.G. Parmigiani, eds. (Springer).
71. Paliy, O., and Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. Mol. Ecol. *25*, 1032–1057.
72. Ramette, A. (2007). Multivariate analyses in microbial ecology. FEMS Microbiol. Ecol. *62*, 142–160.
73. Legendre, P., and Gallagher, E.D. (2001). Ecologically meaningful transformations for ordination of species data. Oecologia *129*, 271–280.
74. Aitchison, J. (1986). The Statistical Analysis of Compositional Data (Springer).
75. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. *8*, 2224.
76. van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., and van der Werf, M.J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genom. *7*, 142.
77. Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol. Direct *4*, 14.
78. Ringnér, M. (2008). What is principal component analysis? Nat. Biotechnol. *26*, 303–304.
79. Hernández Medina, R., Kutuzova, S., Nielsen, K.N., Johansen, J., Hansen, L.H., Nielsen, M., and Rasmussen, S. (2022). Machine learning and deep learning applications in microbiome research. ISME Communications *2*, 1–7.
80. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*.
81. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. Preprint at arXiv. https://doi.org/10.21105/joss.00861.

82. Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance. Austral Ecol. *26*, 32–46.
83. Clarke, K.R. (1993). Non-parametric multivariate analyses of changes in community structure. Austral Ecol. *18*, 117–143.
84. Zuur, A.F., Saveliev, A.A., and Ieno, E.N. (2012). Zero Inflated Models and Generalized Linear Mixed Models with R (Highland Statistics Ltd.).
85. Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A., and Smith, G.M. (2009). Mixed Effects Models and Extensions in Ecology with R (Springer).
86. Wood, S. (2006). Generalized Additive Models: An Introduction with R (Chapman and Hall/CRC).
87. Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., and Hui, F.K.C. (2015). So Many Variables: Joint Modeling in Community Ecology. Trends Ecol. Evol. *30*, 766–779.
88. Chang, H.-X., Haudenshield, J.S., Bowen, C.R., and Hartman, G.L. (2017). Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity. Front. Microbiol. *8*, 519.
89. Grinberg, N.F., Orhobor, O.I., and King, R.D. (2020). An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. Mach. Learn. *109*, 251–277.
90. Lee, S., Kerns, S., Ostrer, H., Rosenstein, B., Deasy, J.O., and Oh, J.H. (2018). Machine Learning on a Genome-wide Association Study to Predict Late Genitourinary Toxicity After Prostate Radiation Therapy. Int. J. Radiat. Oncol. Biol. Phys. *101*, 128–135.
91. Enoma, D.O., Bishung, J., Abiodun, T., Ogunlana, O., and Osamor, V.C. (2022). Machine learning approaches to genome-wide association studies. J. King Saud Univ. Sci. *34*, 101847.
92. Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32.
93. Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. *29*, 1189–1232.
94. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.
95. Feldner-Busztin, D., Firbas Nisantzis, P., Edmunds, S.J., Boza, G., Racimo, F., Gopalakrishnan, S., Limborg, M.T., Lahti, L., and de Polavieja, G.G. (2023). Dealing with dimensionality: the application of machine learning to multi-omics data. Bioinformatics *39*, btad021. https://doi.org/10.1093/bioinformatics/btad021.
96. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.
97. Reel, P.S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. Biotechnol. Adv. *49*, 107739.
98. Pearl, J. (2000). Models, reasoning, and inference. http://library.mpib-berlin.mpg.de/toc/z2008_2219.pdf.
99. Shmueli, G. (2010). To Explain or to Predict? SSO Schweiz. Monatsschr. Zahnheilkd. *25*, 289–310.
100. Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nat. Rev. Genet. *16*, 85–97.
101. Holzinger, E.R., and Ritchie, M.D. (2012). Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. Pharmacogenomics *13*, 213–222.
102. Vellend, M. (2016). The Theory of Ecological Communities (MPB-57) (Princeton University Press).
103. Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., and Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett. *20*, 561–576.
104. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, e8124.
105. Yan, K.K., Zhao, H., and Pang, H. (2017). A comparison of graph- and kernel-based –omics data integration algorithms for classifying complex traits. BMC Bioinf. *18*, 539. https://doi.org/10.1186/s12859-017-1982-4.
106. Holzinger, E.R., Dudek, S.M., Frase, A.T., Pendergrass, S.A., and Ritchie, M.D. (2014). ATHENA: the analysis tool for heritable and environmental network associations. Bioinformatics *30*, 698–705.
107. Tan, K., Huang, W., Hu, J., and Dong, S. (2020). A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. BMC Med. Inform. Decis. Mak. *20*, 129.
108. Garali, I., Adanyeguh, I.M., Ichou, F., Perlbarg, V., Seyer, A., Colsch, B., Moszer, I., Guillemot, V., Durr, A., Mochel, F., and Tenenhaus, A. (2018). A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. Brief. Bioinform. *19*, 1356–1369.
109. Chaudhary, K., Poirion, O.B., Lu, L., and Garmire, L.X. (2018). Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin. Cancer Res. *24*, 1248–1259.
110. Rogozhnikov, A., Ramkumar, P., Bedi, R., Kato, S., and Escola, G.S. (2022). Hierarchical confounder discovery in the experiment-machine learning cycle. Patterns (N Y) *3*, 100451.
111. Hajjem, A., Bellavance, F., and Larocque, D. (2011). Mixed effects regression trees for clustered data. Stat. Probab. Lett. *81*, 451–459.
112. Griffith, G.J., Morris, T.T., Tudball, M.J., Herbert, A., Mancano, G., Pike, L., Sharp, G.C., Sterne, J., Palmer, T.M., Davey Smith, G., et al. (2020). Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat. Commun. *11*, 5749.
113. Arif, S., and MacNeil, M.A. (2022). Predictive models aren't for causal inference. Ecol. Lett. *25*, 1741–1745.
114. Cinelli, C., Forney, A., and Pearl, J. (2022). A crash course in good and bad controls. Sociol. Methods Res. https://doi.org/10.1177/00491241221099552.
115. Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K.B., Vieira, V., Bekker-Jensen, D.B., Kranz, J., Bindels, E.M.J., et al. (2021). Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol. Syst. Biol. *17*, e9730.
116. Mansouri, M., Khakabimamaghani, S., Chindelevitch, L., and Ester, M. (2022). Aristotle: stratified causal discovery for omics data. BMC Bioinf. *23*, 42.