Computational and Structural Biotechnology Journal 19 (2021) 6272–6281

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mathematical-based microbiome analytics for clinical translation

Jayanth Kumar Narayana [a], Micheál Mac Aogáin [b,c], Wilson Wen Bin Goh [a,d], Kelin Xia [e], Krasimira Tsaneva-Atanasova [f], Sanjay H. Chotirmall [a,g,*]

[a] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[b] Biochemical Genetics Laboratory, Department of Biochemistry, St. James's Hospital, Dublin, Ireland
[c] Clinical Biochemistry Unit, School of Medicine, Trinity College Dublin, Dublin, Ireland
[d] School of Biological Sciences, Nanyang Technological University, Singapore, Singapore
[e] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Singapore
[f] Department of Mathematics & Living Systems Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK
[g] Department of Respiratory and Critical Care Medicine, Tan Tock Seng Hospital, Singapore

ARTICLE INFO

Article history:
Received 25 September 2021
Received in revised form 17 November 2021
Accepted 17 November 2021
Available online 22 November 2021

Keywords:
Microbiome
Integration
Mathematical modelling
Microbial association analysis
Topological data analysis
Machine learning

ABSTRACT

Traditionally, human microbiology has been strongly built on the laboratory focused culture of microbes isolated from human specimens in patients with acute or chronic infection. These approaches primarily view human disease through the lens of a single species and its relevant clinical setting however such approaches fail to account for the surrounding environment and wide microbial diversity that exists *in vivo*. Given the emergence of next generation sequencing technologies and advancing bioinformatic pipelines, researchers now have unprecedented capabilities to characterise the human microbiome in terms of its taxonomy, function, antibiotic resistance and even bacteriophages. Despite this, an analysis of microbial communities has largely been restricted to ordination, ecological measures, and discriminant taxa analysis. This is predominantly due to a lack of suitable computational tools to facilitate microbiome analytics. In this review, we first evaluate the key concerns related to the inherent structure of microbiome datasets which include its compositionality and batch effects. We describe the available and emerging analytical techniques including integrative analysis, machine learning, microbial association networks, topological data analysis (TDA) and mathematical modelling. We also present how these methods may translate to clinical settings including tools for implementation. Mathematical based analytics for microbiome analysis represents a promising avenue for clinical translation across a range of acute and chronic disease states.

Contents

* Corresponding author at: Lee Kong Chian School of Medicine, Nanyang Technological University, 11 Mandalay Road, Singapore 308232, Singapore.
E-mail addresses: jayanthk001@e.ntu.edu.sg (J.K. Narayana), m.macaogain@tcd.ie (M. Mac Aogáin), wilsongoh@ntu.edu.sg (W.W.B. Goh), xiakelin@ntu.edu.sg (K. Xia), K.Tsaneva-Atanasova@exeter.ac.uk (K. Tsaneva-Atanasova), schotirmall@ntu.edu.sg (S.H. Chotirmall).

https://doi.org/10.1016/j.csbj.2021.11.029

1. Introduction

1.1. The human microbiome and its role in health and disease

The human microbiome represents a complex and dynamic ecosystem, now established as an important clinical correlate of health and disease [1]. With increasing characterization of the microbiome, our understanding of microbial pathogenesis has progressed exponentially, evolving from focused analysis of individual pathogens to a more holistic analysis incorporating ecological concepts such as diversity, community, and species interaction [2]. The human microbiome is primarily composed of bacteria, viruses and fungi, all of which dynamically interact in a complex manner necessitating multi-dimensional analytic approaches. Such approaches must adapt iteratively as we gain deeper insight into novel aspects of the microbiome's functionality [3–6]. The development of analytical pipelines and mathematical models is therefore critical as it permits deeper exploration of the microbiome towards better clinical insight and potential translation [5,7]. Leveraging upon seminal studies in the gut, human microbiome studies now include multiple anatomic sites and major scientific initiatives such as the human microbiome project have provided a strong base from which the field has evolved [1,3]. Our existence as 'holobionts', composed of human and microbial cells is now clearly established and many physiological processes associate with microbiome composition including digestion, immune regulation and detoxification [1]. Given the increasing appreciation of the centrality of microbes to critical human functions, it is unsurprising that illness and disease is accompanied by significant shifts in microbial composition [8]. Next-generation sequencing (NGS) to derive microbiomes is therefore being increasingly applied across medical disciplines in large observational studies, endophenotyping efforts and clinical trials. Interrogating the generated data requires a careful application of appropriate analytical techniques, and mathematical-based microbiome analytics has emerged as an important means to uncover important signals that may possess potential for clinical translation.

1.2. Next-Generation Sequencing (NGS): Targeted amplicon & metagenomics

Due to the rapid advances in microbial DNA sequencing, our understanding of the human microbiome has rapidly shifted from a microbe-centric, culture-based approach (requiring *a priori* assumptions about the type of sample or disease under investigation) to a less biased pathogen-agnostic NGS approach using a variety of sequencing techniques. Most studies employ targeted amplicon sequencing of the 16S ribosomal RNA gene, one that allows a taxonomic identification of bacteria [9] (Fig. 1A). By corollary, targeting conserved elements of the fungal ribosome, the internally transcribed spacer region (ITS), the fungal microbiome (the mycobiome) can be evaluated, while the analysis of the virome has been challenging and less well defined [10,11]. The current microbiome literature remains heavily biased toward characterisation of the bacteriome, although there is an increasing awareness of the significance for both fungal and viral kingdoms in determining overall composition and function, including the potential for intra-kingdom interaction necessitating integrated analysis [6,10–12]. Whole-genome shotgun (WGS) metagenomics provides a less biased and more holistic alternative to targeted amplicon sequencing and is being increasingly employed in microbiome research (Fig. 1A). While less biased and less susceptible to PCR-associated background contamination, metagenomics is capable of functional profiling although does provide challenges for low biomass samples or samples containing high levels of background human DNA (Fig. 1B). Lower abundance organisms including fungi may be underrepresented or potentially undetected by metagenomics despite serving important biological roles. Here, application of tailored sample preparation methods or blended WGS-target amplicon sequencing approaches may be required to accurately capture true microbial composition. Pacific Biosciences Single Molecule Real Time (SMRT) and Nanopore sequencing protocols further represent ongoing areas of research which promise to bring improvements including full length 16S rRNA gene sequencing and strain-level genomic comparisons but await application in large-scale clinical studies [13–15]. A further important consideration is the analysis of RNA over DNA (metatranscriptomics) which better reflects metabolically active microbes and identifies RNA viruses [16,17]. Notwithstanding such considerations, and the potential pitfalls inherent to the data engineering phase (reviewed in depth elsewhere [9]), the main output of microbiome sequencing workflows remains a set of individual compositional microbiome profiles that serve as a starting point for downstream analysis (Fig. 1B).

1.3. The compositional challenge of microbiome data

In 1897, a classic paper by Karl Pearson indicated the dangers of computing correlations between ratios ($\frac{x_i}{y}$) with common denominators [18]. This paper implied that correlation analysis of compositions $\left(\frac{x_1}{y}, \frac{x_2}{y}, \cdots, \frac{x_n}{y}\right)$, where the ratios ($\frac{x_i}{y}$) are subjected to a sum-constraint $\sum_i^n \frac{x_i}{y} = constant$; may lead to spurious correlations when there are actually none. Components of a composition are

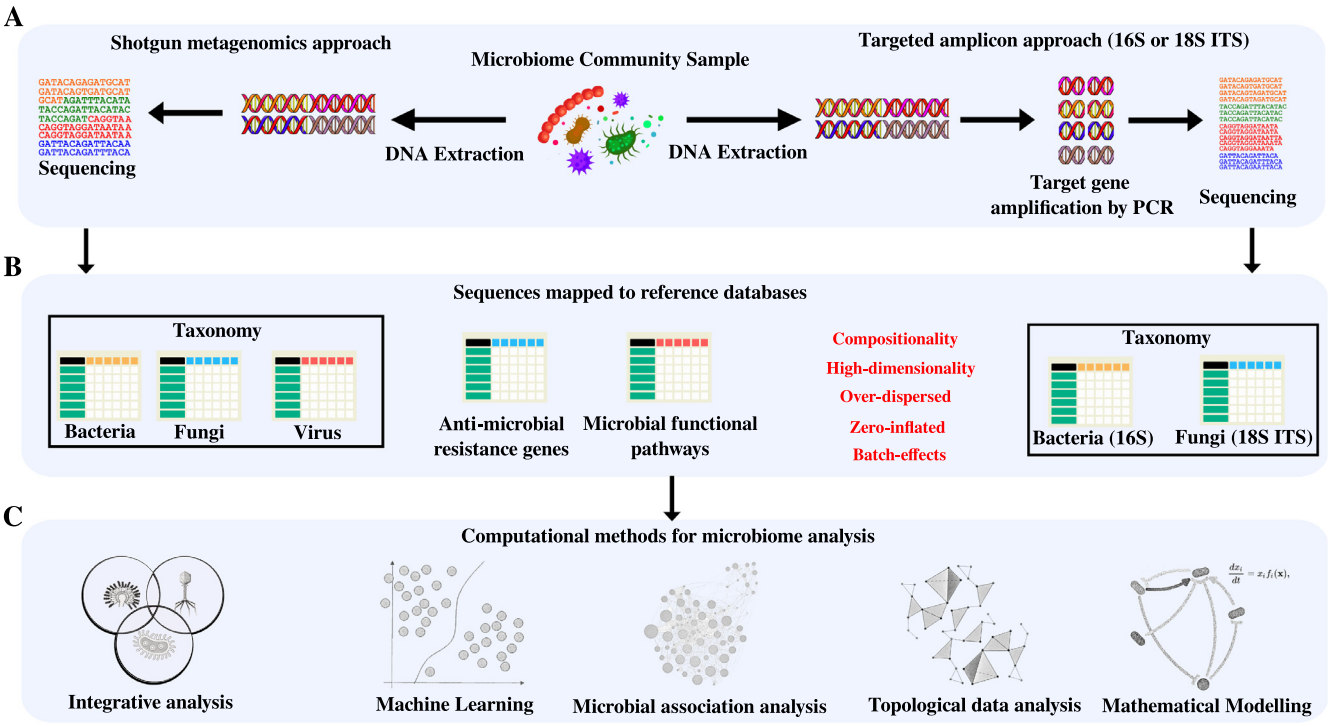

Fig. 1. Overview of the analytical approaches to microbiome data. (A) Microbiome community samples can be assessed by (1) whole genome shotgun metagenomics: where the whole DNA content is sequenced or (2) Targeted amplicon sequencing: where a targeted region (i.e. 16S in bacteria or ITS in fungi) is amplified by polymerase chain reaction (PCR) followed by sequencing. (B) The derived sequences are next mapped to reference databases to yield taxonomic, anti-microbial resistance or functional profiles of the microbiome (whole genome shotgun metagenomics) or taxonomic profile (targeted amplicon sequencing). Derived microbiome profiles suffer from compositionality, high-dimensionality, over-dispersion, sparsity, and batch effects. (C) Various computational approaches for microbiome analytics can be leveraged including integrative microbiome analysis, machine learning, microbial association analysis, topological data analysis and mathematical modelling.

called its parts, and due to its sum-constraint, these samples exist in a mathematical simplex i.e. a 3-part composition resides in a triangle, 4-part in a tetrahedron and so on for higher-dimensional simplexes (see Section 2.4). For example, the relative abundance of n microbial species in a biological sample resides in a n-dimensional simplex space as opposed to an n-dimensional Euclidean space. Most statistical models assume independence between features which does not hold true for compositional data due to its inherent dependency between features. Conventional multivariate data analysis techniques including the Pearson/Spearman correlations, Euclidean distance and multivariate comparisons were developed for data that reside in a Euclidean space and hence not applicable to compositional data. Historically, confusion surrounds compositional data analysis and improper statistical methods have been applied [19]. To address this problem, the Compositional Data Analysis (CoDA) framework was initiated by Aitchison in the 1980s and was based on the theory of log-ratios [19]. This framework has now further extended to include rigorous statistical approaches to analyse compositional datasets [20].

Microbiome datasets derived from next-generation sequencing (NGS) inherit its technical and analytical limitations. In particular, the issue of "normalization" requires consideration given constraints of sequencing capacity i.e., the total number of read counts in a single NGS run. Further, beyond a certain point, the number of reads (or read depth) obtained is irrelevant as it is derived from a random sample of size-selected DNA fragments bound to a sequencing flow cell in accordance with their relative molarity, thus rendering microbiome datasets compositional. Furthermore, each sample does not usually contain exactly the same number of sequence reads and this is attributed to differences in sequencing platforms, experimental difficulties in loading the exact molar amounts of the sequencing libraries and random variation. Hence, microbiome datasets only contain information on their underlying proportions and are often represented as relative abundance, normalized read counts or rarified prior to analysis. Rarefaction involves subsampling of the obtained read counts to a common read depth, however its use is questioned as it leads to loss of potentially useful information [21]. Apart from the statistical complications of microbiome datasets due to their compositionality, the problem is further compounded by high dimensionality and sparsity, as the microbiome comprises several types of microbes and large zero values (Fig. 1B. Hence, studies that use traditional methods to normalize microbiome data, rather than CoDA based methods, may miss important clinical insight due to suboptimal data normalization protocol. For instance, the centered log ratio (CLR) transformation is often used in CoDA based analysis for microbiomes. Given the count vector of 'D' taxa's in a sample $\boldsymbol{x} = [x_1, x_2, \cdots, x_D]$, the CLR transformation of the sample is defined as, $x_{clr} = \left[log\left(\frac{x_1}{G(\boldsymbol{x})}\right), log\left(\frac{x_2}{G(\boldsymbol{x})}\right), \cdots, log\left(\frac{x_D}{G(\boldsymbol{x})}\right)\right]$, where $G(\boldsymbol{x})$ is the geometric mean of x. The clr- transformed values are scale-invariant, i.e. the same ratio is expected to be obtained in a sample with few read counts or with many read counts, only the precision of the estimate is affected [20].

Compositional methods themselves however also suffer from loss-of-scale issues, and therefore efforts toward absolute quantification such as 'spike in' approaches have been employed. This involves the addition of exogenous pre-determined microbial material, that act as an internal control, to derive the absolute abundance of a microbe by 'back-normalization' which mitigates the compositionality of microbiome datasets. These methods however remain under-utilized and largely dependent on the exogeneous additive material. Interestingly, recent work describes the use of a cell-based multi-kingdom spike-in method (MK-

SpikeSeq) to derive absolute abundance, and applies conventional mathematical modelling techniques (see Section 2.5) to derive precise community dynamics of the microbial ecosystem, otherwise not possible using compositional datasets [22].

1.4. Batch effects and the microbiome

Batch effects represent unwanted variation in data caused by factors unrelated to the one of interest, for instance variable experiment times, handlers and reagent lots [23]. Such unwanted effects are endemic in high-throughput methods such as NGS that remain limited by factors such as sequencer capacity, multiple handlers, sample collection, storage or bioinformatic pipelines. Correcting for such effects is imperative, as they otherwise obscure true biological phenomena, reduce statistical power, reproducibility, generalizability or even potentially create artefactual effects [23]. Several Batch Effect Correction Algorithms (BECA) exist but their effectiveness is poorly understood and if inappropriately implemented may lead to loss of biological variation and inflate false positive and false negative rates [24].

Currently, BECAs for microbiome datasets are under-developed and largely derived from gene-expression analysis [25,26]. Batch effects in microbiome data are usually managed by first transforming the dataset using log transformation approaches such as the Centered Log Ratio (CLR) to account for compositionality and sparsity, followed by standard batch correction methods such as ComBat, Batch Mean Centering (BMC) and Surrogate Variable Analysis (SVA), if their assumptions are satisfied [25]. Specifically, for case-control microbiome studies, model-free percentile normalization methods may be implemented for batch effect correction [26]. Limitations to the currently available microbiome batch correction strategies include the often erroneous assumption that a data transformation alone will satisfy the strong assumptions of batch correction methodologies [25]. Additionally, batch-correcting datasets with different microbial community proportions and other imbalances may result in a misestimation of batch-associated variances [27]. To avoid errors due to misassumptions of data normality, non-parametric or distribution-free methods with the ability to accommodate microbiome data characteristics are required. Ideally, such methods must be both effective and precise in targeting batch effects while preserving biological variation [28]. Furthermore, they should be able to cope with batch effects across microbiome studies without being restricted to one type of experimental design (e.g., case–control studies for percentile normalization). Importantly, batch effect-resistant methods such as Similarity Network Fusion (SNF), which mitigate batch effects between -omic datasets (including microbiomes) by creating -omic specific similarity networks prior to merging (see Section 2.1.1) will likely become increasingly relevant in future work and should be considered when developing new analytical methodologies for microbiome analysis.

2. Emerging computational methods for microbiome analytics

2.1. Integrative analysis

Owing to the rapid progress in NGS, we are now able to identify a pool of microbes from human specimens and characterise their taxonomical, functional and resistome profiles [29] (Fig. 1B. Beyond bacteria, fungi, viruses, and their corresponding bacteriophages all represent important components of the human microbiome however most work to date has largely focused on bacteriomes [29]. The key reasons for such bias include a lack of strong reference databases for the viral and fungal kingdoms but also a lack of integrative strategies and computational pipelines to holistically assess intra-kingdom microbiomes. In addition, microbes rarely exist in isolation and usually form complex, interactive, interkingdom communities that encompass the human holobiont [8]. Consequently, a holistic integrative 'multi-biome' approach is most appropriate to accurately represent the true physiological *in vivo* state and gain a greater understanding of any underlying disease pathology.

The development of integrative microbiome analytics has been slow and its progress largely dependent on established integrative -omics methodologies as applied to genomic, transcriptomic, proteomic, epigenomic and metabolomic data [6]. Although microbiome datasets are comparable to other omics, the integration of multi-biome datasets using appropriate analytical methods must be carefully considered before implementation, as such analyses may be influenced by artefacts of the 'omics' technologies itself. The following sections discuss multi-omics integration methodologies that may be applied to microbiome data.

2.1.1. Similarity network Fusion (SNF)

Similarity Network Fusion (SNF) is a network-based multi-omic data integration method that has been successfully applied to microbiomes [12]. For each respective omic dataset, SNF first creates a similarity network using an appropriate measure of similarity. This is next followed by normalization of cross-network similarity scores for individual datasets before merging to create an integrated network that can be applied clinically. Integrated 'patient networks' can then be assessed to classify or identify clinically relevant subgroups based holistically on integrated 'multi-omic' data [6]. SNF provides increased cluster robustness and accuracy, down-weights 'noise' and increases statistical power to detect rarer subgroups from relatively small cohorts. It handles heterogeneous and missing data well however has limitations including its assumption of equal weights (to each integrated dataset) and use of a single similarity metric to capture what is likely complex biological phenomena. Several more recent SNF based methods have been developed to address the various limitations of traditional SNF [6]. Recent work from our group has developed and applied weighted Similarity Network Fusion (wSNF), a method that allows 'weightage' of each individual dataset used in the integration process. We applied wSNF to integrate bacterial, viral and fungal microbiomes in bronchiectasis which resulted in the identification of clinically relevant 'high-risk' patient groups with increased precision as compared to use of single kingdom microbiome datasets [12]. This work serves to underscore the advantage of microbiome data integration in deriving clinically meaningful insights as opposed to single kingdom views of the microbiome.

2.1.2. Data integration analysis for Biomarker discovery using latent cOmponents (DIABLO)

DIABLO is a supervised multi-omics integration strategy developed as part of the mixOmics framework [30]**.** This supervised integrative approach is based on the Generalized Canonical Correlation Analysis that aims to maximize correlations between low-dimensional projections of the input multi-omics datasets. DIABLO improves this by accounting for sparsity of the omics' dataset, a feature particularly relevant to microbiome data and additionally supervises these low-dimensional projections to explain categorical outcomes of interest. Specifically, DIABLO employs discriminant analysis to identify co-expressed or co-related omics features across datasets that may explain the outcome of interest. Importantly, DIABLO assumes linearity, provides no information on causality, and captures only linear relationships which may not hold true in the context of microbiomes. DIABLO also cannot account directly for batch effects that arise across included datasets due to inherent differences in the experimental platforms and analytical pipelines used for each respective dataset [31].

DIABLO has been successfully applied to integrate gut microbiomes with metabolomics, clinical data and microbial function, and this increases classification accuracy compared to singular data analysis, serving to further highlight the analytical gains derived from data integration approaches [32].

2.1.3. Multi-Omics Factor analysis (MOFA)

Multi-Omics Factor Analysis (MOFA) represents an unsupervised statistical framework for multi-omics data integration including microbiomes [33]. MOFA can be considered a generalization of Principal Component Analysis (PCA) that assesses multiple omics datasets as the primary input, with the aim to identify common latent low-dimensional representation of the data. MOFA captures common variation across the various datasets and highlights contributions through feature weights. Important limitations of MOFA include the inability to appropriately capture non-linear relationships and the assumption of independence between features, which in particular may not hold true for microbiomes as microbes exists in communities [34]. MOFA has been employed to integrate bacterial, viral, and fungal components of the intestinal microbiome in critically ill patients with and without sepsis, prior to and following antibiotic exposure. These analyses reveal a modulation of gut microbiomes that involve interkingdom interactions, where overgrowth of potentially invasive viral and fungal organisms is driven by changes to the bacteriome [33].

2.2. Machine learning and microbiomes

Machine Learning (ML) is a class of algorithms that mimic human learning by detecting patterns in data. ML algorithms predict and make decisions without being explicitly programmed but instead use the patterns learnt from the "training data". The use and development of ML algorithms is rapidly accelerating with emerging applications across multiple domains including the practice of medicine. ML methodologies can be categorized into two schemes: (1) supervised and (2) unsupervised. Supervised algorithms learn patterns that map the presented input data to the desired output, given by 'labelled data', using *a priori* identified features for classification and/or regression. In contrast, unsupervised algorithms are only presented with input data and no labelled data, leaving it to find patterns or discover groups in the input dataset which includes clustering. Therefore, a general ML workflow involves (1) data preparation, (2) feature selection, (3) choosing a ML model, (4) training the model, (5) model evaluation, (6) parameter tuning and (7) model testing. As applied to microbiome datasets, ML techniques can have far-reaching benefits for host-trait or disease prediction including risk stratification [35]. The use of ML methods for microbiome-based prediction and classification has been investigated in several studies, with some developing their own framework including DeepMicro, MetaML, Phy-PMFRI, mAML and PopPhy-CNN [36–39]. Machine learning based models have been extensively applied to large-scale microbiome studies with varied applications including the prediction of habitual diet, disease sub-phenotyping and in the identification of 'enterotypes' through clustering and bio-marker identification, for instance linking visceral fat to gut microbial composition [40–42].

2.2.1. DeepMicro

DeepMicro represents a ML framework for microbiome-based prediction. This framework first represents high-dimensional microbiome datasets into low-dimensional latent space followed by implementation of ML models. DeepMicro provides algorithms for dimension reduction such as principal component analysis (PCA), random projection and 'Auto Encoders': a class of unsupervised deep learning algorithms capable of learning representation in terms of latent features and with the potential to capture non-linearities [36]. Following this, DeepMicro provides the user options to implement ML algorithms such Random Forest, Support Vector Machine and Multi-Layer Perceptron algorithms for predictions based on the learned latent representation of the dataset. This framework also employs a thorough k-fold cross-validation scheme for hyper-parameter optimization. It has been established (using five different diseased microbiome datasets) that DeepMicro outperforms other ML frameworks for disease prediction, while significantly reducing dimensionality and training time by 8- to 30-fold suggesting value for microbiome analytics [36].

Critically, however, there are at present only few ML-based microbiome studies that have produced direct clinical translation [35,43]. There are several reasons to explain this: a lack of quality data or sufficient sample size, inadvertent fitting of confounders, lack of generalizable models (i.e. the model does not work beyond the training data) or explainability (i.e. the model cannot be mined for clinical insight), an under-representation of healthy individuals (or non-representativeness in the training set) and heterogeneity of disease phenotypes (hidden sub-labels are present and may confound proper learning), all of which are further complicated by the compositionality and high dimensionality of the underlying datasets. Such limitations may be partially addressed by implementing appropriate data preparatory steps to account for compositionality, batch effects and confounders [44,45]. Further, intelligent feature engineering by combining other techniques such as microbial association networks, mathematical modelling or topological data analysis all hold promise in providing better resolution and robustness for these ML models. Importantly, there remains an unmet need for explainable ML models that support researchers understanding of the underlying biological phenomenon rather than the current provision of only accurately predicting disease classes [46]. Novel and emerging methods including alignment-free (i.e. taxonomy free) approaches that can be leveraged from metagenomic datasets are being increasingly explored and will likely be a focus of future ML applications in the translational space avoiding the bias of current taxonomy-driven approaches [47,48].

2.3. Microbial association analysis

Most biological conditions cannot be attributed to an individual organism but rather a specific microbial signature, consisting of multiple microbes that illustrate the complex underlying ecology that includes microbial interactions [12,22,33]. Network science is therefore recognized as an important technique, for the analysis of complex systems such as that of human microbiomes. Many fundamental discoveries and applications utilize network theory at their core. This includes the Google search algorithm, the discovery of emergent phase transition in material science and the invention of reference models for the internet [49]. In the context of human microbiomes, network science is used to construct microbial association networks [50]. Increasingly, a significant number of methods are being developed for improved microbial network analysis, where microbial clades are represented as nodes and edges (between them) determine associations [51]. A summary of several emerging methods for microbial network inference are detailed below.

2.3.1. Co-occurrence network analysis including renormalization and bootstrap (CoNet)

CoNet is an ensemble-based network inference algorithm producing a weighted undirected graph developed specifically for microbiome datasets [52]. The algorithm uses multiple similarity measures such spearman, kendall and pearson correlations, Kullback–Leibler divergence and bray-curtis similarity to identify consensus microbial similarity networks. Importantly, spurious correlations due to compositionality is accounted for by bootstrap-

ping and renormalization approaches (ReBoot) [53]. The algorithm is available as a Cytoscape plugin and therefore lends itself to implementation for individuals familiar with the Cytoscape suite of functions [52]. CoNet has successfully identified interactions between microbial species in interstitial fibrosis, and is also among pathobionts associated with cancer cachexia [54,55]. Implementation of CoNet with general boosted linear models offers directionality to these microbial associations hinting at causal direction, however, does not establish accurate causality and/or causative mechanisms [53].

2.3.2. Sparse inverse covariance estimation for ecological association inference (SPIEC-EASI)

Spiec-Easi leverages the concept of conditional independence to identify underlying undirected microbial networks by assuming sparsity and estimating the invertible covariance matrix. It employs data transformations developed for compositional data to account for compositionality. Graph creation can be performed by two methods: sparse neighbourhood (which creates a graph node-by-node) and inverse covariance selection (that creates the entire graph) [56]. Finally, stability-based model selection is conducted to infer optimal sparsity of the derived microbial network [56]. Spiec-Easi has been successfully applied to identify cross-kingdom interactions between bacteria and fungi, and to identify key species and assess the topological properties of microbial association networks in antibiotic treated mice [57,58].

2.3.3. Microbial dynamic systems inference engine (MDSINE)

MDSINE is a suite of algorithms using microbiome time series data to infer dynamic system models, which in turn, extrapolate the directed weighted microbial association networks and associated perturbation effect networks [59]. The algorithm takes in two inputs: a temporal microbial abundance and temporal microbial biomass represented for instance by universal 16S rRNA quantitative PCR. This technique accounts for compositionality by estimating non-compositional microbial growth concentrations and their temporal changes using the input datasets. Derived data is then used to infer the parameters of the dynamic system, such as generalized Lotka Volterra models (section 2.5.1). These parameters are then used to infer microbial association networks [59]. MDSINE, in murine *Clostridium difficile* infection models, has been used to determine causal interactions, microbial dynamics, predict stable subcommunities, and to identify microbes most crucial to the integrity of the community under perturbation [59].

Identifying causal relationships between different taxa, including those between the microbiome and other -omics, such as metabolites, remains crucial for understanding the biological mechanisms underlying host-microbe interactions. Clinical translation of microbiome analysis therefore requires the biological mechanisms of interaction between host and microbe to be well understood, essential to make reliable predictions medically and to explore appropriate intervention strategies that have a microbial focus. Experimental methods are therefore most optimal, effective and accurate for establishing such causal relationships [51]. Nevertheless, network inference methodology that generates directed networks using longitudinal datasets such as MDSINE, LSA (Local Similarity Analysis) and TIME (Temporal Insights into Microbial Ecology) can reliably infer causal relationships between taxa [51]. Furthermore, mathematical modelling techniques (Section 2.5) such as genome-scale metabolic modelling along with flux balance analysis may reveal causal interactions between microbes and their hosts [60,61]. These techniques coupled to multi-omics may be used to examine the dynamics of the modelled system in addition to inferring relationships between taxa and their omic-features [62–64]. Advanced analytical methods such as these may then be used to shortlist for instance key relationships for subsequent experimental validation.

Important limitations of such network analysis however include dependence of microbial association networks on sampling resolution, limitations of sequence read analysis and the inability to distinguish live, dead, or dormant cells, although the latter may be accounted for by use of RNA sequencing techniques [50]. Concerns with the interpretability of the output networks remain an important challenge for those using such analysis [50].

2.4. Topological data analysis (TDA) models for microbiomes

Assessment of key microbial taxa is performed by assessing their interaction with other microbes within a community. Microbial association analysis results in microbial networks of interacting taxa (or nodes). Studying and quantifying these structures and associated patterns of the network using graph measurements remains important to identify key components (taxa) of the microbial network. Graph measurements such as degree, stress centrality and betweenness centrality have all been used and shown to have potential clinical translatability and significance [12,65]. In view of the potential clinical utility observed from studying network structures, the field of computational topology, concerned with the study of shape and connectivity, holds significant potential to advance current microbiome analytics.

Topological data analysis (TDA), a widely-used dimensionality reduction and featurization approach is used to study the "shape of data" using models and methods from computational, combinatorial and algebraic topology [66]. Since high dimensional data cannot be directly visualized, TDA may be leveraged to infer topological aspects. TDA methodologies are therefore based on (1) data representation with topological models, including simplicial complexes and hypergraphs and (2) data characterization with topological invariants, including Betti number and Euler characteristics among others. A key feature of TDA is to generalize graphs and networks to simplicial complexes. Physically, edges in networks (or graphs) characterize pair-wise interactions, while complexes, key components of a simplicial complex, characterize higher-dimensional interactions, such as many-body interactions (Fig. 1C). Given the success of network theory in its application to medicine and biology, simplicial complexes and their measures represent a promising avenue for translational research and clinical application. Another key concept in TDA analysis remains topological-invariant, which is a topological measurement that is invariant under continuous deformation. Topological invariant measures represent the most intrinsic and fundamental properties of structures. Among all topological invariants, the most commonly used is the Betti number, which ranks homology groups. Geometrically, dimensionally different Betti numbers represent different homology generators, including their connected components such as loops, circles and holes [66]. Persistent homology is therefore a key model in TDA, providing a bridge between geometry and topology, and in study of the "birth" and "death" of homology generators during a filtration process. Mathematically, the persistence of homology generators encode the geometric information of structures. TDA may derive new insights and deepen our current understanding of microbiome data although the only TDA algorithm applied thus far to microbiome research is Mapper, an approach that aims to uncover and visualize the topological properties of microbiomes [67–70].

Mapper is a computational method for extracting simple descriptions of high dimensional datasets in the form of simplicial complexes. Mapper transforms high-dimensional microbiome profiles into simplicial complexes that reflects geometric aspects i.e., microbiome variation across samples. Network nodes represent a set of samples with similar microbiome profiles and links describe

the intersection of samples between two or more nodes. Mapper accomplishes this in three distinct steps: (1) dimension reduction, (2) covering, and (3) clustering. Dimension reduction is performed by projecting high-dimensional data points (representing microbiome profiles) to a low-dimensional space using user-defined functions called 'filter' functions. Filter functions can also be referred to as 'lenses' because the selection of such functions potentially reveal different aspects of the dataset. For example, Shannon diversity and Berger-Parker dominance indices act as 'filter' functions that represent the sample based on its taxonomic diversity and taxonomic dominance as views, which are not necessarily the same. Upon reducing dimension, Mapper divides this low-dimensional space into several covers of equal size overlapping with one another. Covers in this space capture local neighbourhoods of data, and the overlap connects these neighbourhoods to capture global structure. Finally, Mapper implements clustering on the pull-back of each cover, to group samples with similar microbiome profiles. This step is critical to retain the original distance information of the high-dimensional microbiome profile as information concerning original distances between samples may be potentially lost after dimension reduction. For example, two samples that are far apart in high-dimensional space (i.e. dissimilar in microbiome profile) might be projected as close neighbours in the low-dimensional space (e.g., due to similar diversity values). The simplicial complex is then generated such that each node represents a cluster, and a link is then drawn between nodes if they share common samples within their clusters.

The adoption of TDA based techniques (such as Mapper) to microbiome datasets has now led to frameworks such as tmap which recently illustrates superiority in detecting non-linearities in data, for instance in enterotype analysis, driver species identification, and microbiome-wide association analysis in conjunction host metadata [70]. Mapper has been successfully used to uncover state transitions in human gut microbiomes and asthma endotypes based on microbiome profiles [67–69]. Mapper importantly outperforms PCA and PCoA in distinguishing patient characteristics based on microbiome profiles [71].

2.5. Mathematical modelling of the microbiome

Most microbial association analyses can identify significant dependency between microbes within the microbiome however such dependencies cannot predict 'causality' and a system's future behaviour. Despite this, the identification of such dependency may be used to build dynamic mathematical models of the microbial community that may be used to assess causality, i.e. the effects of its different components on one another and make predictions about its behaviour. In addition, mathematical models are constructed based on the understanding and/or assumptions of the system's mechanisms, hence, are invaluable in studying a system's behaviour to changes in its parameters and/or validating assumed mechanisms through experimentation. The power of explainability; inherent to mathematical models can also be used as a complementary strategy along with data analysis, that aims to extract information from data for clinical application. However, thus far, few studies have attempted to apply such modelling techniques to microbial communities [62]. Within a mathematical model, the modelling units themselves represent the most basic interacting entities of the overall system; for instance, taxa (or microbial species), individual cells, functional guilds, or the overall community. The choice of unit dictates model resolution, the subsequent simulated dynamics and defines the potential frameworks that may be employed for analysis. These include the following frameworks: (1) supra-organismal; (2) population-based; (3) heterogenous and/or (4) integrative [62].

2.5.1. Lotka Volterra models

The commonest modelling approach employed in microbiome research remains population-based models such as the generalized Lotka Volterra Models (gLV). These methods study population dynamics of the modelling units and assumes homogeneity of internal states across the individual microbes within each population. Such pairwise models describe potential relationships between 'n' species (or taxa) using ordinary differential equations to track their population growth dynamics i.e.:

$$\frac{ds_i}{dt} = s_i\left(\mu_i + \sum_{j=1}^{n} a_{ij}x_j\right), j = 1, \cdots, n$$

where s_i represents species (or taxon) abundance, 'μ_i' its growth rate and 'a_{ij}' the interaction strength between species (or taxa) *'i'* and *'j'*. Although this model is widely used and relatively easy to implement, it (1) requires an absolute quantification of abundance to account for compositionality, (2) assumes the additive influence of fitness from pairwise interactions, and (3) does not provide any understanding of chemical intermediates [72]. Alternative modelling techniques of finer model resolution may partially address some of these limitations such as mechanistic modelling, that offers insight into chemical intermediaries between microbes, stochiometric modelling which offers mechanistic insight and compositional Lotka-Volterra models, that allows modelling on compositional data [62,73]. In addition, thermodynamic models, evolutionary game theory models and integrative modelling strategies may all be potentially useful in clinical microbiomics as they each integrate various other modelling strategies of different resolutions thereby potentially overcoming the inherent weaknesses when only a single approach is used [62].

Generalized Lotka Volterra models have been used to predict gut microbial dynamics in preterm infants accounting for environmental perturbations and revealing microbial blooms [22]. Jones et al. applied gLV models in antibiotic-induced *C. difficile* infection to evaluate microbial dynamics and proposes mathematical model-motivated experiments [74]. Mathematical modelling techniques including community metabolic modelling have been leveraged to derive metabolic networks in the gut and when applied to type 2 diabetes reveals unique network structures that render a significant metabolic influence *in vivo* [75]. To successfully apply mathematical modelling to microbiome analytics, the appropriate choice of model to use is key. Selection should be driven by the intended objective and follow *Occam's razor* principle i.e. to use the simplest credible model given the biological question under investigation. Such models are characterised by fewest parameters thereby reducing the complexity of the model structure and hence the probability of over-fitting. Available tools such as Akiake Information Criterion (AIC) and Bayesian Information Criterion (BIC) can also systematically and quantitatively compare between models. The mathematical modelling of microbiomes is likely to be an area of significant development for potential clinical translation over the coming years as it allows an understanding of microbial associations under specific conditions, predicts dynamics under fresh conditions and models the outcomes when microbial communities are controlled to perform clinically advantageous functions.

3. Summary and outlook

In this review, various emerging mathematical approaches to microbiome analytics have been outlined while emphasizing their potential application in clinical translation (Table 1). Each approach has its strengths however general challenges inherent to microbiome data need to be acknowledged and addressed including compositionality and batch effects. While community

Table 1
Table summarizing open-source tools and software available for the methods described in this manuscript.

Method	Tools/Softwares available
Compositional Data Analysis	'compositions' – a R package [76] 'propr' – a R package [77] 'CoDa Pack' – a multiplatform standalone software [78]
Similarity Network Fusion (SNF)	'SNFtool' – a R package [79] 'Integrative Microbiomics' – a webtool for integration of microbiomes [80]
Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO)	Part of 'mixOmics' – a R package [81]
Multi-Omics Factor Analysis (MOFA)	'MOFA2' – a R package [82]
DeepMicro	'DeepMicro' – a python package [83]
Co-occurrence network analysis including renormalization and bootstrap (CoNet)	'CoNet' – a cytoscape app [84]
Sparse inverse covariance estimation for ecological association inference (SPIEC-EASI)	'SpeicEasi' – a R package [85]
Microbial dynamic systems inference engine (MDSINE)	'mdsine' – available as standalone and MATLAB library [86]
Mapper	'Kepler Mapper' – a python package [87]

ecological techniques including ordination (Principal Coordinate Analysis), diversity, dominance analysis and discriminant taxa analysis such as Indicator species analysis, Similarity Percentages (SIMPER) and Linear discriminant analysis Effect Size (LEfSe) are all commonly employed in microbiome research, a clear need for more advanced models and analytical techniques is now required to match the complexity and rapid development of available NGS approaches. Additionally, current microbiome studies are largely observational and descriptive in nature and there is a clear need for rigorous mathematical and analytical approaches to attain a more precise understanding of the role of microbiomes in human disease. Realizing the underlying biological mechanisms of microbial dynamics, dysbiosis and its interaction with the host is critical for clinical translatability. Considering this, novel yet rigorous mathematical approaches will be necessary to meet the demands to better understand, shortlist 'causal relationships' and subsequently engineer microbial communities to impact clinical medicine. It is likeliest that a combination of the presented techniques, applied in an appropriate setting holds significant potential for clinical translation, however, such analytical approaches must be validated in rigorous experimental models to provide confidence in practical applications that may include the engineering of individual species to manipulate a community or as microbial biomarkers in specific disease-states.

CRediT authorship contribution statement

Jayanth Kumar Narayana: Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Micheál Mac Aogáin:** Conceptualization, Methodology, Writing – review & editing. **Wilson Wen Bin Goh:** Conceptualization, Methodology, Writing – review & editing. **Kelin Xia:** Writing – review & editing. **Krasimira Tsaneva-Atanasova:** Conceptualization, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Sanjay H. Chotirmall:** Conceptualization, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the Singapore Ministry of Health's National Medical Research Council under its Clinician-Scientist Individual Research Grant (CS-IRG) (MOH-000141) (S.H.C) and Clinician Scientist Award (CSA) (MOH-000710) (S.H.C). It is also supported by the NTU Integrated Medical, Biological and Environmental Life Sciences (NIMBELS), Nanyang Technological University, Singapore [NIM/03/2018] (S.H.C). KTA gratefully acknowledges the financial support of the EPSRC via grant EP/T017856/1. The authors would like to acknowledge The Academic Respiratory Initiative for Pulmonary Health (TARIPH) for collaboration support. S.H.C declares membership of advisory boards and speaker fees from Astra-Zeneca, CSL Behring and Boehringer-Ingelheim all outside of the submitted work. All other authors have no conflicts of interest to disclose.

References

[1] Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med 2018;24:392–400. https://doi.org/10.1038/nm.4517.
[2] Gilbert JA, Lynch SV. Community ecology as a framework for human microbiome research. Nat Med 2019;25:884–9. https://doi.org/10.1038/s41591-019-0464-9.
[3] Gevers D, Pop M, Schloss PD, Huttenhower C, Eisen JA. Bioinformatics for the Human Microbiome Project. PLoS Comput Biol 2012;8:e1002779. https://doi.org/10.1371/journal.pcbi.1002779.
[4] Morgan XC, Huttenhower C. Chapter 12: Human microbiome analysis. PLoS Comput Biol 2012;8:e1002808. 10.1371/journal.pcbi.1002808.
[5] Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes Dis 2017;4:138–48. https://doi.org/10.1016/j.gendis.2017.06.001.
[6] Narayana JK, Aogáin MM, Ali NABM, Tsaneva-Atanasova K, Chotirmall SH. Similarity network fusion (SNF) for the integration of multi-omics and microbiomes in respiratory disease. Eur Respir J 2021;2101016. https://doi.org/10.1183/13993003.01016-2021.
[7] Bucci V, Xavier JB. Towards predictive models of the human gut microbiome. J Mol Biol 2014;426:3907–16. https://doi.org/10.1016/j.jmb.2014.03.017.
[8] Pitlik SD, Koren O. How holobionts get sick-toward a unifying scheme of disease. Microbiome 2017;5:64. https://doi.org/10.1186/s40168-017-0281-7.
[9] Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform 2021;22:178–93. https://doi.org/10.1093/bib/bbz155.
[10] Tiew PY, Mac Aogain M, Ali NABM, Thng KX, Goh K, Lau KJX, et al. The mycobiome in health and disease: emerging concepts, methodologies and challenges. Mycopathologia 2020;185:207–31. https://doi.org/10.1007/s11046-019-00413-z.
[11] Liang G, Bushman FD. The human virome: assembly, composition and host interactions. Nat Rev Microbiol 2021;19:514–27. https://doi.org/10.1038/s41579-021-00536-5.
[12] Mac Aogáin M, Narayana JK, Tiew PY, Ali NABM, Yong VFL, Jaggi TK, et al. Integrative microbiomics in bronchiectasis exacerbations. Nat Med 2021;27:688–99. https://doi.org/10.1038/s41591-021-01289-7.
[13] Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, et al. Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. Microbiome 2018;6. https://doi.org/10.1186/s40168-018-0569-2.
[14] Maghini DG, Moss EL, Vance SE, Bhatt AS. Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. Nat Protoc 2021;16:458–71. https://doi.org/10.1038/s41596-020-00424-x.
[15] Aogáin MM, Jaggi TK, Chotirmall SH. The airway microbiome: present and future applications. Arch Bronconeumol 2021. https://doi.org/10.1016/j.arbres.2021.08.003.

[16] Sulaiman I, Wu BG, Li Y, Tsay J-C, Sauthoff M, Scott AS, et al. Functional lower airways genomic profiling of the microbiome to capture active microbial metabolism. Eur Respir J 2021;58:2003434. https://doi.org/10.1183/13993003.03434-2020.
[17] Wylie KM. The virome of the human respiratory tract. Clin Chest Med 2017;38:11–9. https://doi.org/10.1016/j.ccm.2016.11.001.
[18] Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond 1897;60:489–98. 10.1098/rspl.1896.0076.
[19] Aitchison J, editor. The statistical analysis of compositional data. Dordrecht: Springer Netherlands; 1986.
[20] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol 2017;8:2224. https://doi.org/10.3389/fmicb.2017.02224.
[21] McMurdie PJ, Holmes S, McHardy AC. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 2014;10:e1003531. https://doi.org/10.1371/journal.pcbi.1003531.
[22] Rao C, Coyte KZ, Bainter W, Geha RS, Martin CR, Rakoff-Nahoum S. Multi-kingdom ecological drivers of microbiota assembly in preterm infants. Nature 2021;591:633–8. https://doi.org/10.1038/s41586-021-03241-8.
[23] Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol 2017;35:498–507. https://doi.org/10.1016/j.tibtech.2017.02.012.
[24] Goh WWB, Wong L. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects –- a case study in clinical proteomics. BMC Genomics 2017;18:142. https://doi.org/10.1186/s12864-017-3490-3.
[25] Wang Y, LêCao K-A. Managing batch effects in microbiome data. Brief Bioinf 2020;21:1954–70. https://doi.org/10.1093/bib/bbz105.
[26] Gibbons SM, Duvallet C, Alm EJ, Langille M. Correcting for batch effects in case-control microbiome studies. PLoS Comput Biol 2018;14:e1006102. https://doi.org/10.1371/journal.pcbi.1006102.
[27] Zhou L, Chi-Hau Sue A, Bin Goh WW. Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects? J Genet Genom 2019;46:433–43. https://doi.org/10.1016/j.jgg.2019.08.002.
[28] Wang Y, Lê Cao K-A. A multivariate method to correct for batch effects in microbiome data. Bioinformatics 2020. https://doi.org/10.1101/2020.10.27.358283.
[29] Pflughoeft KJ, Versalovic J. Human microbiome in health and disease. Annu Rev Pathol Mech Dis 2012;7:99–122. https://doi.org/10.1146/annurev-pathol-011811-132421.
[30] Rohart F, Gautier B, Singh A, Lê Cao K-A, Schneidman D. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 2017;13:e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
[31] Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics 2019;35:3055–62. 10.1093/bioinformatics/bty1054.
[32] Meslier V, Laiola M, Roager HM, De Filippis F, Roume H, Quinquis B, et al. Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. Gut 2020;69:1258–68. https://doi.org/10.1136/gutjnl-2019-320438.
[33] Haak BW, Argelaguet R, Kinsella CM, Kullberg RFJ, Lankelma JM, Deijs M, et al. Integrative transkingdom analysis of the gut microbiome in antibiotic perturbation and critical illness. MSystems 2021;6. 10.1128/mSystems.01148-20.
[34] Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 2020;21. https://doi.org/10.1186/s13059-020-02015-1.
[35] Sakowski E, Uritskiy G, Cooper R, Gomes M, McLaren MR, Meisel JS, et al. Current state of and future opportunities for prediction in microbiome research: Report from the Mid-Atlantic Microbiome Meet-up in Baltimore on 9 January 2019. MSystems 2019;4. 10.1128/mSystems.00392-19.
[36] Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep 2020;10:6026. https://doi.org/10.1038/s41598-020-63159-5.
[37] Pasolli E, Truong DT, Malik F, Waldron L, Segata N, Eisen JA. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 2016;12:e1004977. https://doi.org/10.1371/journal.pcbi.1004977.
[38] LaPierre N, Ju C-J-T, Zhou G, MetaPheno WW. A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. Methods 2019;166:74–82. https://doi.org/10.1016/j.ymeth.2019.03.003.
[39] Yang F, Zou Q. mAML: an automated machine learning pipeline with a microbiome repository for human disease classification. Database 2020;2020:baaa050. https://doi.org/10.1093/database/baaa050.
[40] Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat Med 2021;27:321–32. https://doi.org/10.1038/s41591-020-01183-8.
[41] MetaHIT Consortium (additional members), Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. Enterotypes of the human gut microbiome. Nature 2011;473:174–80. 10.1038/nature09944.
[42] He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 2018;24:1532–5. https://doi.org/10.1038/s41591-018-0164-x.
[43] Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nat Rev Gastroenterol Hepatol 2020;17:635–48. https://doi.org/10.1038/s41575-020-0327-3.
[44] Wong L. Big data and a bewildered lay analyst. Stat Prob Lett 2018;136:73–7. https://doi.org/10.1016/j.spl.2018.02.033.
[45] Goh WWB, Wong L. Dealing with confounders in omics analysis. Trends Biotechnol 2018;36:488–98. https://doi.org/10.1016/j.tibtech.2018.01.013.
[46] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.
[47] Ding X, Cheng F, Cao C, Sun X. DectICO: an alignment-free supervised metagenomic classification method based on feature extraction and dynamic selection. BMC Bioinf 2015;16:323. https://doi.org/10.1186/s12859-015-0753-3.
[48] Asgari E, Garakani K, McHardy AC, Mofrad MRK. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. Bioinformatics 2018;34:i32–42. https://doi.org/10.1093/bioinformatics/bty296.
[49] Gosak M, Markovič R, Dolenšek J, Slak Rupnik M, Marhl M, Stožer A, et al. Network science of biological systems at different scales: A review. Phys Life Rev 2018;24:118–35. https://doi.org/10.1016/j.plrev.2017.11.003.
[50] Faust K. Open challenges for microbial network construction and analysis. ISME J 2021;15:3111–8. https://doi.org/10.1038/s41396-021-01027-4.
[51] Dohlman AB, Shen X. Mapping the microbial interactome: Statistical and experimental approaches for microbiome network inference. Exp Biol Med (Maywood) 2019;244:445–58. https://doi.org/10.1177/1535370219836771.
[52] Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape. F1000Res 2016;5:1519. https://doi.org/10.12688/f1000research.
[53] Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol 2012;8:e1002606. https://doi.org/10.1371/journal.pcbi.1002606.
[54] Jacob N, Jacobs JP, Kumagai K, Ha CWY, Kanazawa Y, Lagishetty V, et al. Inflammation-independent TL1A-mediated intestinal fibrosis is dependent on the gut microbiome. Mucosal Immunol 2018;11:1466–76. https://doi.org/10.1038/s41385-018-0055-y.
[55] Pötgens SA, Brossel H, Sboarina M, Catry E, Cani PD, Neyrinck AM, et al. Klebsiella oxytoca expands in cancer cachexia and acts as a gut pathobiont contributing to intestinal dysfunction. Sci Rep 2018;8. https://doi.org/10.1038/s41598-018-30569-5.
[56] Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA, et al. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol 2015;11:e1004226. https://doi.org/10.1371/journal.pcbi.1004226.
[57] Tipton L, Müller CL, Kurtz ZD, Huang L, Kleerup E, Morris A, et al. Fungi stabilize connectivity in the lung and skin microbial ecosystems. Microbiome 2018;6. https://doi.org/10.1186/s40168-017-0393-0.
[58] Mahana D, Trent CM, Kurtz ZD, Bokulich NA, Battaglia T, Chung J, et al. Antibiotic perturbation of the murine gut microbiome enhances the adiposity, insulin resistance, and liver disease associated with high-fat diet. Genome Med 2016;8. https://doi.org/10.1186/s13073-016-0297-9.
[59] Bucci V, Tzen B, Li N, Simmons M, Tanoue T, Bogart E, et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. Genome Biol 2016;17. https://doi.org/10.1186/s13059-016-0980-6.
[60] Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol 2010;28:245–8. https://doi.org/10.1038/nbt.1614.
[61] Eisenstein M. Microbial matters: modelling the complex web of host–microbiome interactions. Nature 2020;581:489–90. https://doi.org/10.1038/d41586-020-01472-9.
[62] Song H-S, Cannon W, Beliaev A, Konopka A. Mathematical modeling of microbial community dynamics: A methodological review. Processes 2014;2:711–52. https://doi.org/10.3390/pr2040711.
[63] Coyte KZ, Rao C, Rakoff-Nahoum S, Foster KR. Ecological rules for the assembly of microbiome communities. PLoS Biol 2021;19:. https://doi.org/10.1371/journal.pbio.3001116e3001116.
[64] Gonze D, Coyte KZ, Lahti L, Faust K. Microbial communities as dynamical systems. Curr Opin Microbiol 2018;44:41–9. https://doi.org/10.1016/j.mib.2018.07.004.
[65] Quinn RA, Whiteson K, Lim YW, Zhao J, Conrad D, LiPuma JJ, et al. Ecological networking of cystic fibrosis lung infections. NPJ Biofilms Microbiomes 2016;2:4. https://doi.org/10.1038/s41522-016-0002-1.
[66] Wasserman L. Topological data analysis. Annu Rev Stat Appl 2018;5:501–32. https://doi.org/10.1146/annurev-statistics-031017-100045.
[67] Abdel-Aziz MI, Brinkman P, Vijverberg SJH, Neerincx AH, Riley JH, Bates S, et al. Sputum microbiome profiles identify severe asthma phenotypes of relative stability at 12 to 18 months. J Allergy Clin Immunol 2021;147:123–34. https://doi.org/10.1016/j.jaci.2020.04.018.
[68] Hinks T, Zhou X, Staples K, Dimitrov B, Manta A, Petrossian T, et al. Multidimensional endotypes of asthma: topological data analysis of cross-sectional clinical, pathological, and immunological data. The Lancet 2015;385:S42. https://doi.org/10.1016/S0140-6736(15)60357-9.

[69] Chang WK, VanInsberghe D, Kelly L. Topological analysis reveals state transitions in human gut and marine bacterial communities. NPJ Biofilms Microbiomes 2020;6:41. https://doi.org/10.1038/s41522-020-00145-9.
[70] Liao T, Wei Y, Luo M, Zhao G-P, Zhou H. tmap: an integrative framework based on topological data analysis for population-scale microbiome stratification and association studies. Genome Biol 2019;20:293. https://doi.org/10.1186/s13059-019-1871-4.
[71] Yazdani M, Smarr L, Knight R. Using Topological Data Analysis to find discrimination between microbial states in human microbiome data, 2016.
[72] Momeni B, Xie L, Shou W. Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. ELife 2017;6:. https://doi.org/10.7554/eLife.25051e25051.
[73] Joseph TA, Shenhav L, Xavier JB, Halperin E, Pe'er I. Compositional Lotka-Volterra describes microbial dynamics in the simplex. PLoS Comput Biol 2020;16:e1007917. 10.1371/journal.pcbi.1007917.
[74] Jones EW, Carlson JM. In silico analysis of antibiotic-induced Clostridium difficile infection: Remediation techniques and biological adaptations. PLoS Comput Biol 2018;14:. https://doi.org/10.1371/journal.pcbi.1006001e1006001.
[75] Sung J, Kim S, Cabatbat JJT, Jang S, Jin Y-S, Jung GY, et al. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. Nat Commun 2017;8:15393. https://doi.org/10.1038/ncomms15393.
[76] Boogaart KG van den, Tolosana-Delgado R, Bren M. compositions: compositional data analysis. 2021.
[77] Quinn T, Lovell D, Erb I, Bilgrau A, Gloor G, Moore R. propr: Calculating proportionality between vectors of compositional data. 2019.
[78] Comas-Cufí M, Thió-Henestrosa S. CoDaPack 2011. http://ima.udg.edu/codapack/ (accessed November 1, 2021).
[79] Wang B, Mezlini A, Demir F, Fiume M, Tu Z, Brudno M, et al. SNFtool: Similarity Network Fusion. 2021.
[80] Narayana JK, Mac Aogáin M, Tsaneva-Atanasova K, Chotirmall SH. Integrative Microbiomics n.d. https://integrative-microbiomics.ntu.edu.sg/ (accessed November 1, 2021).
[81] Cao K-AL, Rohart F, Gonzalez I, Dejean S, Abadi AJ, Gautier B, et al. mixOmics: Omics Data Integration Project. Bioconductor version: Release (3.14); 2021. 10.18129/B9.bioc.mixOmics.
[82] Argelaguet R, Arnol D, Bredikhin D, Velten B. MOFA2: Multi-Omics Factor Analysis v2. Bioconductor version: Release (3.14); 2021. 10.18129/B9.bioc.MOFA2.
[83] minoh0201. DeepMicro. 2021.
[84] Cytoscape App Store - CoNet n.d. https://apps.cytoscape.org/apps/conet (accessed November 2, 2021).
[85] Kurtz Z. SpiecEasi. 2021.
[86] MDSINE / mdsine — Bitbucket n.d. https://bitbucket.org/MDSINE/mdsine/src/master/ (accessed November 2, 2021).
[87] van Veen H, Saul N, Eargle D, Mangham S. Kepler Mapper: A flexible Python implementation of the Mapper algorithm. JOSS 2019;4:1315. 10.21105/joss.01315.