The diagnostic potential and barriers of microbiome based therapeutics

Acharjee, Animesh; Singh, Utpreksha; Choudhury, Saptamita Paul; Gkoutos, Georgios

DOI:
10.1515/dx-2022-0052

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Acharjee, A, Singh, U, Choudhury, SP & Gkoutos, G 2022, 'The diagnostic potential and barriers of microbiome based therapeutics', *Diagnosis*. https://doi.org/10.1515/dx-2022-0052

Opinion Paper

Animesh Acharjee*, Utpreksha Singh, Saptamita Paul Choudhury and Georgios V. Gkoutos

The diagnostic potential and barriers of microbiome based therapeutics

https://doi.org/10.1515/dx-2022-0052
Received May 28, 2022; accepted August 3, 2022;
published online August 25, 2022

Abstract: High throughput technological innovations in the past decade have accelerated research into the trillions of commensal microbes in the gut. The 'omics' technologies used for microbiome analysis are constantly evolving, and large-scale datasets are being produced. Despite of the fact that much of the research is still in its early stages, specific microbial signatures have been associated with the promotion of cancer, as well as other diseases such as inflammatory bowel disease, neurogenerative diareses etc. It has been also reported that the diversity of the gut microbiome influences the safety and efficacy of medicines. The availability and declining sequencing costs has rendered the employment of RNA-based diagnostics more common in the microbiome field necessitating improved data-analytical techniques so as to fully exploit all the resulting rich biological datasets, while accounting for their unique characteristics, such as their compositional nature as well their heterogeneity and sparsity. As a result, the gut microbiome is increasingly being demonstrating as an important component of personalised medicine since it not only plays a role in inter-individual variability in health and disease, but it also represents a potentially modifiable entity or feature that may be addressed by treatments in a personalised way. In this context, machine learning and artificial intelligence-based methods may be able to unveil new insights into biomedical analyses through the generation of models that may be used to predict category labels, and continuous values. Furthermore, diagnostic aspects will add value in the identification of the non invasive markers in the critical diseases like cancer.

Keywords: biomarker; diagnostics; machine learning; microbiota.

***Corresponding author: Dr. Animesh Acharjee**, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK; Institute of Translational Medicine, University of Birmingham, Birmingham, UK; NIHR Surgical Reconstruction and Microbiology Research Centre, University Hospital Birmingham, B15 2TT, Birmingham, UK; and MRC Health Data Research UK (HDR UK), Midlands Site, Birmingham, UK, Phone: +4407403642022, E-mail: a.acharjee@bham.ac.uk. https://orcid.org/0000-0003-2735-7010
Utpreksha Singh, Department of Health and Life Sciences, Coventry University, Coventry, UK
Saptamita Paul Choudhury, KIIT School of Biotechnology, KIIT University, Bhubaneswar, India
Georgios V. Gkoutos, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK; Institute of Translational Medicine, University of Birmingham, Birmingham, UK; NIHR Surgical Reconstruction and Microbiology Research Centre, University Hospital Birmingham, Birmingham, UK; MRC Health Data Research UK (HDR UK), Birmingham, UK; and NIHR Experimental Cancer Medicine Centre, Birmingham, UK

Background

Technological innovations in the past decade have accelerated research into the trillions of commensal microbes in the gut, the microbiome. Across gut, inflammatory, and neurological disorders, there are alterations in gut microbiome composition and function [1, 2]. As a result, a plethora of machine learning approaches are increasingly being developed targeting the microbiome due to its diagnostic potential [3]. While the majority of these machine learning studies are critically flawed [4], the microbiome remains a potentially invaluable source for identifying novel biomarkers [5].

Culture-based diagnostics for monitoring

A simple, albeit limited, diagnostic approach involves selectively-culturing fecal samples to monitor antibiotic-resistant bacteria, such as *Enterococci* [6]. However, this approach is limited to well-identified organisms that are easily cultured [5] and have been successfully demonstrated to aid the diagnosis of gastrointestinal infections [7, 8]. For many exploratory studies, culture-based methods are far too limited in their scope but could provide additional functional information that can form valuable components for developing novel machine learning-based analysis approaches.

16S rRNA diagnostics

The availability and declining cost of sequencing has rendered RNA-based diagnostics more common in the microbiome field. Amplifying and sequencing conserved bacterial ribosomal regions results into microbiome diversity and composition high resolution data [9, 10]. Such datasets have the potential of aiding prediction approach to depict patients that are likely to develop bacterial-resistant infections; for example, *Morganella* and *Prevotella* abundance conferred resistance [7]. 16S rRNA sequencing offers the tantalising possibility of catering the development of models that cater the prediction subsequent infection with antibiotic-resistant bacteria [11].

Beyond gastrointestinal disorders, several studies suggest that 16S microbiome biomarkers may be useful for differentiating Alzheimer's from dementia, as well as detecting schizophrenia, Parkinson's disease, inflammatory bowel disease, colorectal and other cancers, osteoporosis, bipolar disorder, and other immune-related diseases [12–22]. Other applications include predicting Crohn's disease relapse and drug responses [23, 24].

Metagenomic diagnostics

Metagenomic sequencing provides a more precise microbiome data resolution which can detect fungal and viral commensals [9, 25]. In addition to the increasingly accurate species-level resolution, it also detects specific genes that are being expressed in the microbiome, providing functional information [26]. Some studies leverage this extra layer of information to generate diagnostics, though it isn't always clear whether they provide a cost-effective advantage.

Thus far, several studies have employed metagenomic features or variables generated from experiments to diagnose early-stage colorectal cancer, different stages of Parkinson's disease, fatty-liver disease, schizophrenia, as well as other brain and gut disorders [27–31]. Importantly, future diagnostic metagenomic based approach may be able to facilitate the early detection of neurodegenerative diseases or depict particular patient drug-responses within psychiatric settings [32].

Metabolomic diagnostics

Fecal metabolomics provides more functional information that can be used to detect microbiome metabolites, such as short-chain fatty acids, tryptophan metabolites and secondary bile acids [33]. Since it is used less frequently than other methods, it has been employed by fewer diagnostic applications compared to those utilising 16S or metagenomic data. Nevertheless, metabolomics datasets have been proven useful for differentiating healthy controls from patient cohorts suffering from certain cancers, myasthenia gravis, lupus, cardiovascular disease, liver disease, and other disorders [29, 34–37].

Outlook

The high number of features or variables within the microbiome, as well as its key role in physiology renders it an ideal source of diagnostic biomarkers across multiple diseases. Most studies select differentially abundant microbial features that are applied within diagnostic models. Predominantly, these microbial features are detected through 16S, whole genome metagenomics, or metabolomics. However, many studies still lack consistency and replicability and have yet to be validated in clinical trial settings.

Emergent problems in microbiome datasets

Both sequencing, as well as metabolomic, microbiome datasets are compositional nature [38–40]. Simply put, sequencing technology imposes an arbitrary detection limit of individual elements. Similarly, to a forest or other ecosystems, the abundance of one species or metabolite affects the abundance of other elements – prohibiting these elements to be treated as independent variables.

It follows that each detected element is present within this interdependent composition, rendering them more challenging to interpret and analyze [38, 41]. Thus we cannot consider this data as true counts of elements in the gut environment, but instead we need to contemplate them as a proportional count of molecules associated with microbes in the gut [38].

Normalization to account for sparsity and zeroes

The first step in analyzing any raw counts table involves normalizing the values it contains to allow for their comparison. However, typically, upon inspecting such data, it becomes obvious that it is sparse, and unevenly distributed. Moreover, such data usually contains several zeroes which need to be properly interpreted as true zeros or artefacts [38, 42].

Running robust multivariate comparisons involves generating Monte Carlo samples of the Dirichlet distribution within each sample before normalization [39, 43]. While many studies use rarefaction subsampling methods, these, typically, result in information loss across the overall composition [44]. Although, Ddta can be transformed using simple log-ratios, such an approach does not sufficiently address the sparsity and zeroes within the dataset [39, 40]. Tools, such as ALDEx2, allow users to perform central, isometric, or additive log ratios for normalization that conserves the relationships between elements better [39, 43].

Distance and ordination

Typically machine learning discrimination models attempt to separate two or more independent groups based on their unique features or variables. However, many common forms of ordination, typically employed in microbiome data analysis, employ non-compositional methods, such as UniFrac, Bray-Curtis, and Jensen-Shannon divergence [38, 41]. The Aitchison's distance is a more reliable method for measuring distance or volatility of compositions over time, providing a geometric measure of distance between principal components [41, 45].

Differential abundance and effect size

Following a multivariate statistics analysis, the abundance and effect size differences need to be determined. While 95% confidence intervals and effect size cut-offs may help identify salient features within a dataset, there is no specific cut-off to indicate clinical or real-world relevance [33]. It is unclear if it is necessary for some microbes or metabolites to reach a certain threshold before exerting strong effects. Small changes in other important microbes by contrast, could destabilize a microbial ecosystem if these microbes serve as important hubs within the network [46–50].

Correlation, causality and directionality

The compositional nature of microbiome datasets dictate that any correlations and comparisons are susceptible to negative correlation bias [38, 51, 52], an affect that necessitated the development of methods, such as SparCC and SpiecEasi [38]. While there aren't any robust methods for determining causality and directionality, most research methods are now focusing on proportionality i.e., how much of a particular phenotype is explained by a feature within the dataset [53–59]. While Granger causality has also been suggested as a potential method for microbiome analysis, it is seldom used and yet to be validated [58]. It should finally be noted that whether any amount of data or any method development will ever suffice to determine causality is still questionable [60].

Outlook

The compositional nature of microbiome data imposes several challenges and biases that renders their analysis challenging. In addition to the sparsity and zeroes within the data, methods need to account for negative correlation bias, as well as many other uncertainties relating to the importance of different effect sizes.

Machine learning methods and workflows

There are many different types of machine learning methods, broadly split into supervised and unsupervised learning. In supervised learning, data is fully labelled whereas in unsupervised clustering, no data is labelled. Different types of semi-supervised approaches classify a small portion of the dataset for use as a test set.

Unsupervised clustering

Unsupervised clustering methods, such as K-means, have been implemented to characterize microbial features across stages of periodontitis [61]. Principal component analysis is a valuable approach to cluster compositional data, measuring distances based on the Aitchison's metric to determine whether there is a significant difference between two different groups of samples [38, 39, 41]. Non-negative matrix factorization and t-distributed stochastic neighbor embedding (t-SNE) are seldom used with microbiome datasets – though they may be appropriate for describing some cohorts [62].

For multifactorial disease datasets, such as depression related data, unsupervised clustering might provide powerful means for feature-based stratification that could indicate the best course of treatment.

Supervised learning

Supervised learning methods are far more common across microbiome studies and are especially useful in identifying associations between host phenotypes and microbial features [62]. These methods are used for classification or regression analyses.

Regression

Lasso, ridge, and elastic net are different types of penalized regression. These are especially useful when there are many more features within the data than sample size. The penalty function is important to prevent overinflation within the regression model [62]. The L1 penalty is used in lasso regression which shrinks the size of coefficients while setting the values of coefficients for unimportant features to zero; additionally, if two variables show collinearity, only one of these two variables will have a non-zero coefficient [63, 64].

In ridge regression, the L2 square root function penalty is used to shrink values near zero though it does not remove these variables [65]. Since the variables are included as non-zero coefficients, their effects are incorporated into the machine learning model. Meanwhile the lasso net method is a linear combination of lasso and ridge, retaining all features as distinct groups of variables [62, 66].

In the literature, regression methods are used across different microbiome studies to identify brain, immune, diabetic, and gastrointestinal disorders as well as cancers based on a host phenotype [19, 28, 67–71].

Classification

The support vector machine (SVM) approach employs linear or non-linear distances and margins to segregate different groups [72]. Linear applications use L1 and L2 penalty functions along with the SVM, thereby retrieving the most salient differentiating features and determining the maximal boundary between multiple groups, a task not performed by simple L1 or L2 regression. A comparison to common regression methods demonstrated that microbiome models using various SVMs, although outperforming random forest methods, don't always outperform simpler L2 regressions [73]. However, a scoping review of different machine learning techniques revealed an opposing trend, with random forest performing well against other types of models [74].

The random forest approach uses a re-sampling method, termed bootstrapping, along with random selection, to generate different types of classifier decision trees [75]. It determines the variable importance of different features from these decision trees, allowing users to select the most salient metabolites or genes. When used appropriately, these models can detect features of colorectal cancer [76]. These preselected important features are also useful in future mechanistic studies.

Another popular method used within the microbiome literature involves linear discriminant analysis, part of the LEfSe package [77]. It is similar to a supervised version of Principal Component analysis albeit inappropriate for compositional datasets [38].

Artificial neural networks and deep learning

Advances over the last decade allowed for bioinformatics approaches to employ advanced approaches, such as artificial neural networks. These networks relay an input layer of information, an activation function that combines the information, and finally an output function [62]. By manipulating the number of computational layers, or how many times data is passed back and forth between input and output, users can increase or decrease the complexity and power of the model. The goal of an artificial neural network is to find the ideal weights for different features to optimize classification [62, 78, 79].

Deep learning models, such as tensor flow, employ many more layers, allowing them to outperform other algorithms when handling large amounts of data [62]. This, however, comes at the expense of model and feature interpretability. Some tools also combine Bayesian prediction with deep learning, to assess whether some known host features, combined with microbiome data, can predict disease [80].

Reinforcement learning

Other machine learning methods employ rules instead of labels to guide the model development [81]. These methods could be harnessed for finding bacterial proteins or metabolites whose structures may fit human protein receptors.

Machine learning and AI workflows

Many python-based tools incorporate genomic data into reproducible workflows. The BioBakery workflow encompasses several different tools for metagenomic analysis developed within the Huttenhower lab; many of these tools

use linear discriminant analysis [82]. The individual programs and components within this workflow can also be adapted to integrate multi-omics data [83]. Microbiome Analyst is another example of a complete workflow albeit one that incorporates several tools that are inappropriate for compositional data analysis [84, 85].

DeepMicro converts high-dimensional data into low-dimensional representations thereby rendering their interpretation and analysis easier [86]. NFnetFu is a new pipeline demonstrated to perform well compared to other commonly used microbiome pipelines and tools, using a novel combination of approaches to process the data [87]. Briefly, an adaptive neuro fuzzy inference system preprocesses the data to overcome sparsity, even across smaller sample sizes, while maintaining the collinearity and relationships between features [87]. Then, a density-based clustering method is applied to reduce the collinearity into a simpler matrix for further analysis, followed by a lasso L1 regression [87]. Finally, it concludes with a taxon set enrichment analysis to identify relevant biological networks within the dataset [87]. A graphical overview is presented in the Figure 1.

Outlook

While there exist numerous different types of machine learning models and bioinformatics applications and frameworks, it is often difficult to select the best one without benchmarking them. Additionally, it is unclear where the increase of the model complexity leads to better predictions or more actionable insights. Among existing machine learning algorithms, linear discriminant analysis is most commonly used across microbiome studies to identify differential features between groups [77].

Fecal microbiota transplantation (FMT) as diagnostic process

Fecal microbiota transplantation (FMT) is the delivery of solutions of fecal material from donors entering the digestive system of a receiver in an attempt to actively affect the patient's microbiological diversity and bestow a therapeutic advantage [88, 89]. FMT originated in Chinese

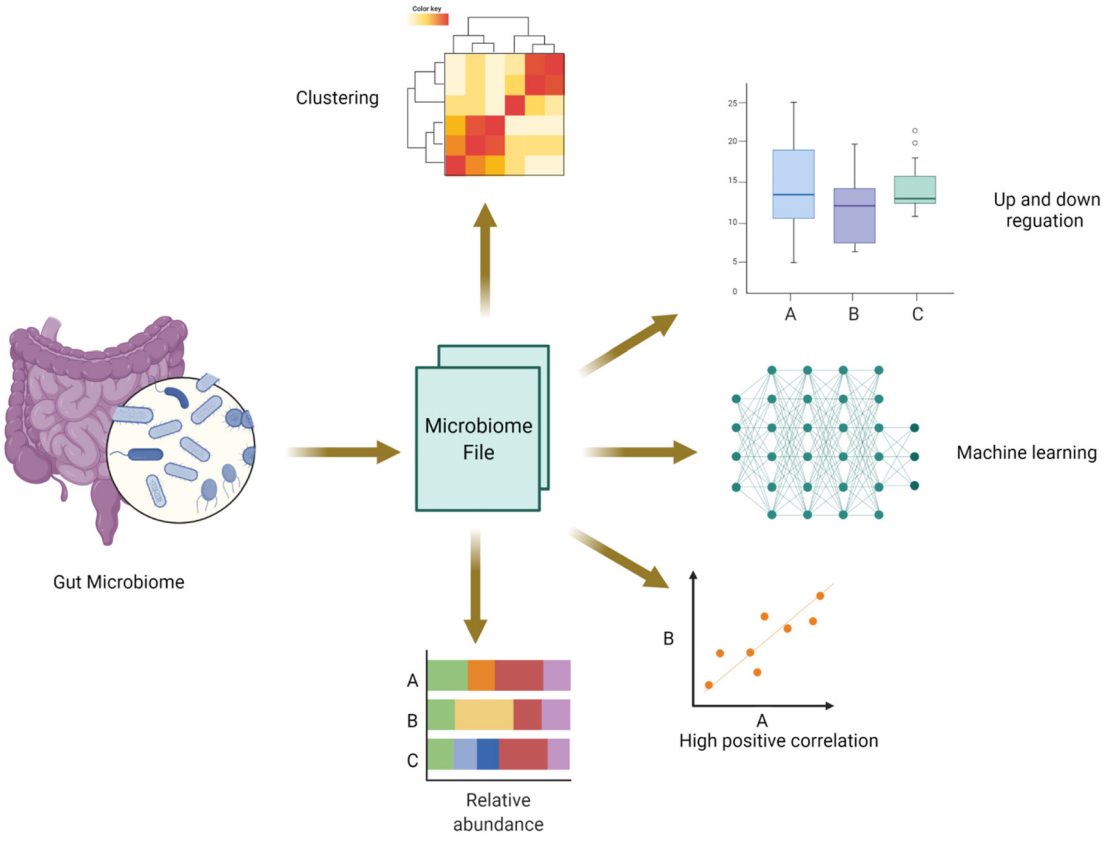

Figure 1: Machine learning-based analysis approaches employing microbiome datasets.

civilization over 1,700 years ago in the 4th century, when a well conventional Chinese medical practitioner entitled Ge Hong effectively cured sufferers of foodborne poisoning and/or chronic diarrhea via oral utilizing a humanoid fecal suspension recognized as "yellow soup" [90, 91]. Dr. Ben Eiseman, Chief of Surgery at Denver General Hospital, along with his team commented on their successful usage of fecal enemas to cure four individuals with pseudomembranous colitis [92]. Scientists didn't recognize that the illness they had been treating, pseudomembranous colitis, has been induced by a virulent bacteria called *Clostridioides difficile* until 1978, 20 years later [93].

Individuals can receive fecal microbiome through a number of methods, involving enema [94], nasogastric or nasoduodenal tubes [95] colonoscopy [96], or oral capsules [97, 98]. Furthermore, it is increasingly normal to use refrigerated or freeze-dried feces via anonymously, pre-screened, fit contributors. Despite other antimicrobial treatments, FMT does not produce an underlining dysbiosis which makes the sufferer vulnerable to infections [99]. The processes that enable FMT to prevent infections are still being studied [100], but they most probably include intestinal ecological restorations, intestinal flora composition and functioning, and microbe-host signaling transmission.

Emerging medical research, including animal model research, is increasingly pointing to the value of FMT in inflammatory bowel disease and metabolic syndrome patient treatment. FMT has also been proposed as a viable therapy for various psychiatric diseases, notably autism spectrum disorder. Although experimental animal models have revealed that FMT is effective in preventing energy metabolism disorder and other disorders linked to the gutmicrobiome these results should be carefully considered [101].

Outlook and challenges

As outlined previously, microbiome data offers immense therapeutic potential but remains challenging to analyse and interpret for a variety of reasons including its compositional structure nature that introduces a negative correlation bias, collinearity and sparsity (i.e., multiple types of zeroes value). Currently, the majority of existing microbiome diagnostic approaches suffer from proper validation which result in leakage or overfitting [4].

Lost in translation

Assume that the comparison between a disease and a control group reveals that a microbe is differentially expressed and quantified – how is its clinical relevance then determined? There are no standard cut-offs for effect sizes nor any requirements for reporting 95% confidence intervals and frequently the confidence interval for these differentially abundant features includes zero [33]. But even if these challenges are addressed there remains the issue of determining the directionality of the microbiome diversity impact, as well as the proportion of the disease phenotype that is attributed to the unique microbial features. Moreover, and perhaps more importantly, there needs to be a systematic framework in place that will enable the testing and validation of the efficacy and hypotheses generated through these studies by other laboratories and/or clinical settings.

Applying macro-ecological approaches

Nonetheless, microbiome researchers continue to adapt and develop classic macro-ecological microbiome approaches. Future applications may include developing machine learning models with multi-label classification capacity [102] as well as capable of taking into account volatility and temporal microbiome changes [45, 103]. Given that there is no "standard" or "healthy" microbiome, it remains a challenge to develop replicable diagnostic microbiome pipelines.

This necessitates the need of including more personalized information accompanied by a description of potential factors impacting ecosystems. Many studies now carefully integrate dietary information along with other types of metadata to define the impact of a patient specific microbiome [3, 50, 81, 104–107]. Unfortunately, it isn't always clear whether the increasing complexity resulting from the application of advanced machine learning approaches, that take into account larger and more diverse datasets, is beneficial and to what degree. Undeniably, in certain contexts, simple linear regression can still outperform more expensive, time-consuming and resource heavy approaches [108].

Fecal microbiota transplantation is now a recognised therapy for Clostridium Difficle Infection, Inflammatory Bowel Disease, as well as for several intestinal disorders [94]. Next-generation sequencing technology improvements

have dramatically increased the scope of microbiome variations that may be defined [109], enabling the discovery of dysbiosis biomarkers, as well as the development of increasingly tailored therapy methods, to enhance the effectiveness or accessibility of FMT. Several majors, as well as linked difficulties, should be solved in an attempt to enhance FMT procedures. Firstly, the definition of a "good health" or "perfectly natural" microbe needs to be clearly determined in an attempt to get a deterministic comprehension of host-microbiome interactions. Then the definition of a "good health" or "regular" microbe needs to be clearly stated in sequence to enable more easy detection as well as adjustment of dysbiotic states. Finally, modifications in gastrointestinal microbiology have to be effectively characterized in the literature Although FMT has clear limitations, it offers the potential of a readily available treatment approach that benefits from reduced toxic effects in comparison to existing synthetic pharmaceutical based approaches.

Outlook

Despite an abundance of recent, novel machine learning approaches and frameworks developed within the context of translational research, very few examples have been demonstrated to successful achieve and materialise such translations in clinical contexts. Each tool has its own advantages and limitations, which need to be considered when selecting a workflow or algorithm. Moreover, it is vital to assess the performance of the workflow against other standard algorithms and workflows. Nevertheless, we anticipate the future development and use of machine learning models as invaluable diagnostic or pre-screening tools.

Research funding: The authors acknowledge support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, Nanocommons H2020-EU (731032), MAESTRIA (Grant agreement ID 965286) and the MRC Heath Data Research UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health.
Author contributions: A.A. conceptualized and designed the strategy. U.S., S.P.C contributes to the literature review and A.A. U.S., S.P.C., G.V.G. wrote the first draft; All authors have accepted responsibility for the entire content of this manu script and approved its submission.
Competing interests: Authors state no conflict of interest.
Informed consent: Not applicable.
Ethical approval: Not applicable.

References

1. Cryan JF, O'Riordan KJ, Cowan CSM, Sandhu KV, Bastiaanssen TFS, Boehme M, et al. The microbiota-gut-brain Axis. Physiol Rev 2019;99:1877–2013.
2. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature 2016;535:94–103.
3. Lavelle A, Hill C. Gut microbiome in health and disease: emerging diagnostic opportunities. Gastroenterol Clin N Am 2019;48:221–35.
4. Quinn TP. Stool studies don't pass the sniff test: a systematic review of human gut microbiome research suggests widespread misuse of machine learning; 2021. New York: arXiv. preprint arXiv:210703611.
5. Damhorst GL, Adelman MW, Woodworth MH, Kraft CS. Current capabilities of gut microbiome-based diagnostics and the promise of clinical application. J Infect Dis 2021;223:(Suppl 3) S270–5.
6. Humphreys H. Controlling the spread of vancomycin-resistant enterococci. Is active screening worthwhile? J Hosp Infect 2014; 88:191–8.
7. Pettigrew MM, Gent JF, Kong Y, Halpin AL, Pineles L, Harris AD, et al. Gastrointestinal microbiota disruption and risk of colonization with carbapenem-resistant Pseudomonas aeruginosa in intensive care unit patients. Clin Infect Dis 2019; 69:604–13.
8. Halpin AL, de Man TJ, Kraft CS, Perry KA, Chan AW, Lieu S, et al. Intestinal microbiome disruption in patients in a long-term acute care hospital: a case for development of microbiome disruption indices to improve infection prevention. Am J Infect Control 2016;44:830–6.
9. Clooney AG, Fouhy F, Sleator RD, AOD, Stanton C, Cotter PD, et al. Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis. PLoS One 2016;11:e0148028.
10. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol 2016;16:123.
11. Taur Y, Xavier JB, Lipuma L, Ubeda C, Goldberg J, Gobourne A, et al. Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. Clin Infect Dis 2012;55:905–14.
12. Zheng P, Yang J, Li Y, Wu J, Liang W, Yin B, et al. Gut microbial signatures can discriminate unipolar from bipolar depression. Adv Sci 2020;7:1902862.
13. Liu P, Wu L, Peng G, Han Y, Tang R, Ge J, et al. Altered microbiomes distinguish Alzheimer's disease from amnestic mild cognitive impairment and health in a Chinese cohort. Brain Behav Immun 2019;80:633–43.

14. Ticinesi A, Nouvenne A, Tana C, Prati B, Meschi T. Gut microbiota and microbiota-related metabolites as possible biomarkers of cognitive aging. Adv Exp Med Biol 2019;1178:129–54.
15. Shen Y, Xu J, Li Z, Huang Y, Yuan Y, Wang J, et al. Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia: a cross-sectional study. Schizophr Res 2018;197:470–7.
16. Xu Z, Xie Z, Sun J, Huang S, Chen Y, Li C, et al. Gut microbiome reveals specific dysbiosis in primary osteoporosis. Front Cell Infect Microbiol 2020;10:160.
17. Sze MA, Schloss PD. Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. mBio 2018;9. https://doi.org/10.1128/mBio.02076-18.
18. Manandhar I, Alimadadi A, Aryal S, Munroe PB, Joe B, Cheng X. Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases. Am J Physiol Gastrointest Liver Physiol 2021. https://doi.org/10.1152/ajpgi.00360.2020.
19. Dong M, Li L, Chen M, Kusalik A, Xu W. Predictive analysis methods for human microbiome data with application to Parkinson's disease. PLoS One 2020;15:e0237779.
20. Kanangat S, Skaljic I. Microbiome analysis, the immune response and transplantation in the era of next generation sequencing. Hum Immunol 2021. https://doi.org/10.1016/j.humimm.2021.07.009.
21. Fukui H, Nishida A, Matsuda S, Kira F, Watanabe S, Kuriyama M, et al. Usefulness of machine learning-based gut microbiome analysis for identifying patients with irritable bowels syndrome. J Clin Med 2020;9:1141.
22. Wipperman MF, Bhattarai SK, Vorkas CK, Maringati VS, Taur Y, Mathurin L, et al. Gastrointestinal microbiota composition predicts peripheral inflammatory state during treatment of human tuberculosis. Nat Commun 2021;12:1141.
23. Sprockett D, Fischer N, Boneh RS, Turner D, Kierkus J, Sladek M, et al. Treatment-specific composition of the gut microbiota is associated with disease remission in a pediatric crohn's disease cohort. Inflamm Bowel Dis 2019;25:1927–38.
24. Doherty MK, Ding T, Koumpouras C, Telesco SE, Monast C, Das A, et al. Fecal microbiota signatures are associated with response to ustekinumab therapy among crohn's disease patients. mBio 2018;9. https://doi.org/10.1128/mBio.02120-17.
25. Claesson MJ, Clooney AG, O'Toole PW. A clinician's guide to microbiome analysis. Nat Rev Gastroenterol Hepatol 2017;14:585–95.
26. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9:811–4.
27. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 2017;66:70–8.
28. Bedarf JR, Hildebrand F, Coelho LP, Sunagawa S, Bahram M, Goeser F, et al. Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naive Parkinson's disease patients. Genome Med 2017;9:39.
29. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, et al. Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. Cell Metabol 2019;30:607.
30. Wang S, Li N, Zou H, Wu M. Gut microbiome-based secondary metabolite biosynthetic gene clusters detection in Parkinson's disease. Neurosci Lett 2019;696:93–8.
31. Zhu F, Ju Y, Wang W, Wang Q, Guo R, Ma Q, et al. Metagenome-wide association of gut microbiome features for schizophrenia. Nat Commun 2020;11:1612.
32. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, et al. The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol 2019;4:623–32.
33. Spichak S, Bastiaanssen TFS, Berding K, Vlckova K, Clarke G, Dinan TG, et al. Mining microbes for mental health: determining the role of microbial metabolic pathways in human brain health and disease. Neurosci Biobehav Rev 2021;125:698–761.
34. Feng J, Zhao F, Sun J, Lin B, Zhao L, Liu Y, et al. Alterations in the gut microbiota and metabolite profiles of thyroid carcinoma patients. Int J Cancer 2019;144:2728–45.
35. Tan X, Huang Y, Chai T, Zhao X, Li Y, Wu J, et al. Differential gut microbiota and fecal metabolites related with the clinical subtypes of myasthenia gravis. Front Microbiol 2020;11:564579.
36. Yan R, Jiang H, Gu S, Feng N, Zhang N, Lv L, et al. Fecal metabolites were altered, identified as biomarkers and correlated with disease activity in patients with systemic lupus erythematosus in a GC-MS-based metabolomics study. Front Immunol 2020;11:2138.
37. Li J, Zuo K, Zhang J, Hu C, Wang P, Jiao J, et al. Shifts in gut microbiome and metabolome are associated with risk of recurrent atrial fibrillation. J Cell Mol Med 2020;24:13356–69.
38. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol 2017;8:2224.
39. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2014;2:15.
40. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One 2013;8:e67019.
41. Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V. Logratio analysis and compositional distance. Math Geol 2000;32:271–5.
42. Prost V, Gazut S, Brüls T. A zero inflated log-normal model for inference of sparse microbial association networks. PLoS Comput Biol 2021;17:e1009089.
43. Quinn TP, Crowley TM, Richardson MF. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinf 2018;19:274.
44. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 2014;10:e1003531.
45. Bastiaanssen TFS, Gururajan A, van de Wouw M, Moloney GM, Ritz NL, Long-Smith CM, et al. Volatility as a concept to understand the impact of stress on the microbiome. Psychoneuroendocrinology 2021;124:105047.

46. 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.
47. Zhang Z, Geng J, Tang X, Fan H, Xu J, Wen X, et al. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. ISME J 2014;8:881–93.
48. Hall CV, Lord A, Betzel R, Zakrzewski M, Simms LA, Zalesky A, et al. Co-Existence of network architectures supporting the human gut microbiome. iScience 2019;22:380–91.
49. Jackson MA, Bonder MJ, Kuncheva Z, Zierer J, Fu JY, Kurilshikov A, et al. Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. PeerJ 2018;6:e4303.
50. Imangaliyev S, Keijser B, Crielaard W, Tsivtsivadze E. Personalized microbial network inference via co-regularized spectral clustering. Methods 2015;83:28–35.
51. Filzmoser P, Walczak B. What can go wrong at the data normalization step for identification of biomarkers? J Chromatogr A 2014;1362:194–205.
52. Filzmoser P, Hron K, Reimann C. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci Total Environ 2009;407:6100–8.
53. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bahler J. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol 2015;11:e1004075.
54. Quinn TP, Erb I. Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data. NAR Genom Bioinform 2020;2:lqaa076.
55. Walter J, Armet AM, Finlay BB, Shanahan F. Establishing or exaggerating causality for the gut microbiome: lessons from human microbiota-associated rodents. Cell 2020;180:221–32.
56. Wade KH, Hall LJ. Improving causality in microbiome research: can human genetic epidemiology help? Wellcome Open Res 2019;4:199.
57. Vorholt JA, Vogel C, Carlstrom CI, Muller DB. Establishing causality: opportunities of synthetic communities for plant microbiome research. Cell Host Microbe 2017;22:142–55.
58. Mainali K, Bewick S, Vecchio-Pagan B, Karig D, Fagan WF. Detecting interaction networks in the human microbiome with conditional Granger causality. PLoS Comput Biol 2019;15: e1007037.
59. Koh A, Bäckhed F. From association to causality: the role of the gut microbiota and its functional products on host metabolism. Mol Cell 2020. https://doi.org/10.1016/j.molcel.2020.03.005.
60. Lynch KE, Parke EC, O'Malley MA. Microbiome causality: further reflections (a response to our commentators). Biol Philos 2020; 35:29.
61. Boutin S, Hagenfeld D, Zimmermann H, El Sayed N, Höpker T, Greiser HK, et al. Clustering of subgingival microbiota reveals microbial disease ecotypes associated with clinical stages of periodontitis in a cross-sectional study. Front Microbiol 2017;8: 340.
62. Namkung J. Machine learning methods for microbiome studies. J Microbiol 2020;58:206–16.
63. Liu Z, Bensmail H, Tan M. Efficient feature selection and multiclass classification with integrated instance and model based learning. Evol Bioinform Online 2012;8:197–205.
64. Chen L, Liu H, Kocher JP, Li H, Chen J. glmgraph: an R package for variable selection and predictive modeling of structured genomic data. Bioinformatics 2015;31:3991–3.
65. Waldron L, Pintilie M, Tsao MS, Shepherd FA, Huttenhower C, Jurisica I. Optimized application of penalized regression methods to diverse genomic data. Bioinformatics 2011;27: 3399–406.
66. Susin A, Wang Y, Lê Cao KA, Calle ML. Variable selection in microbiome compositional data analysis. NAR Genom Bioinform 2020;2:lqaa029.
67. Han L, Zhao K, Li Y, Han H, Zhou L, Ma P, et al. A gut microbiota score predicting acute graft-versus-host disease following myeloablative allogeneic hematopoietic stem cell transplantation. Am J Transplant 2020;20:1014–27.
68. Cai L, Wu H, Li D, Zhou K, Zou F. Type 2 diabetes biomarkers of human gut microbiota selected via iterative sure independent screening method. PLoS One 2015;10:e0140827.
69. Zhang W, Chen X, Wong KC. Noninvasive early diagnosis of intestinal diseases based on artificial intelligence in genomics and microbiome. J Gastroenterol Hepatol 2021;36:823–31.
70. Ye F, Wu X, Wang T, Liang J, Li J, Dai Y, et al. Identification of immune-associated gene signature and immune cell infiltration related to overall survival in progressive multiple sclerosis. Mult Scler Relat Disord 2021;55:103188.
71. Nicholson K, Bjornevik K, Abu-Ali G, Chan J, Cortese M, Dedi B, et al. The human gut microbiota in people with amyotrophic lateral sclerosis. Amyotroph Lateral Scler Frontotemporal Degener 2021;22:186–94.
72. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20:273–97.
73. Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. A framework for effective application of machine learning to microbiome-based classification problems. mBio 2020;11: e00434–20.
74. Iadanza E, Fabbri R, Bašić-ČiČak D, Amedei A, Telalovic JH. Gut microbiota and artificial intelligence approaches: a scoping review. Health Technol 2020;10:1–16.
75. Breiman L. Random forests. Mach Learn 2001;45:5–32.
76. Young C, Wood HM, Seshadri RA, Van Nang P, Vaccaro C, Melendez LC, et al. The colorectal cancer-associated faecal microbiome of developing countries resembles that of developed countries. Genome Med 2021;13:27.
77. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol 2011;12:R60.
78. Reiman D, Farhat AM, Dai Y. Predicting host phenotype based on gut microbiome using a convolutional neural network approach. Methods Mol Biol 2021;2190:249–66.
79. Reiman D, Layden BT, Dai Y. MiMeNet: exploring microbiome-metabolome relationships using neural networks. PLoS Comput Biol 2021;17:e1009021.
80. Sharma D, Xu W. phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data. Bioinformatics 2021;3707–14. https://doi.org/10.1093/bioinformatics/btab482.
81. McCoubrey LE, Elbadawi M, Orlu M, Gaisford S, Basit AW. Harnessing machine learning for development of microbiome therapeutics. Gut Microb 2021;13:1–20.

82. McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, et al. bioBakery: a meta'omic analysis environment. Bioinformatics 2018;34:1235–7.
83. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, Mohney RP, et al. Interplay between the human gut microbiome and host metabolism. Nat Commun 2019;10:4505.
84. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res 2017;45:W180–8.
85. Chong J, Liu P, Zhou G, Xia J. Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. Nat Protoc 2020;15:799–821.
86. Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep 2020;10: 6026.
87. Bisht V, Acharjee A, Gkoutos GV. NFnetFu: a novel workflow for microbiome data fusion. Comput Biol Med 2021;135:104556.
88. Bakken JS, Borody T, Brandt LJ, Brill JV, Demarco DC, Franzos MA, et al. Treating Clostridium difficile infection with fecal microbiota transplantation. Clin Gastroenterol Hepatol 2011;9: 1044–9.
89. Smits LP, Bouter KE, de Vos WM, Borody TJ, Nieuwdorp M. Therapeutic potential of fecal microbiota transplantation. Gastroenterology 2013;145:946–53.
90. Shi Y-C, Yang Y-S. Fecal microbiota transplantation: current status and challenges in China. JGH Open 2018;2:114–6.
91. Chang CW, Lee HC, Li LH, Chiang Chiau JS, Wang TE, Chuang WH, et al. Fecal microbiota transplantation prevents intestinal injury, upregulation of toll-like receptors, and 5-fluorouracil/ oxaliplatin-induced toxicity in colorectal cancer. Int J Mol Sci 2020:21. https://doi.org/10.3390/ijms21020386.
92. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One 2012;7:e49138.
93. George RH, Symonds JM, Dimock F, Brown JD, Arabi Y, Shinagawa N, et al. Identification of Clostridium difficile as a cause of pseudomembranous colitis. Br Med J 1978;1:695.
94. Youngster I, Sauk J, Pindar C, Wilson RG, Kaplan JL, Smith MB, et al. Fecal microbiota transplant for relapsing Clostridium difficile infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. Clin Infect Dis 2014;58:1515–22.
95. Staley C, Khoruts A, Sadowsky MJ. Contemporary applications of fecal microbiota transplantation to treat intestinal diseases in humans. Arch Med Res 2017;48:766–73.
96. Kelly CR, Khoruts A, Staley C, Sadowsky MJ, Abd M, Alani M, et al. Effect of fecal microbiota transplantation on recurrence in multiply recurrent Clostridium difficile infection: a randomized trial. Ann Intern Med 2016;165:609–16.
97. Staley C, Hamilton MJ, Vaughn BP, Graiziger CT, Newman KM, Kabage AJ, et al. Successful resolution of recurrent Clostridium difficile infection using freeze-dried, encapsulated fecal microbiota; pragmatic cohort study. Am J Gastroenterol 2017;112:940–7.
98. Youngster I, Russell GH, Pindar C, Ziv-Baran T, Sauk J, Hohmann EL. Oral, capsulized, frozen fecal microbiota transplantation for relapsing Clostridium difficile infection. JAMA 2014;312:1772–8.
99. Cammarota G, Masucci L, Ianiro G, Bibbò S, Dinoi G, Costamagna G, et al. Randomised clinical trial: faecal microbiota transplantation by colonoscopy vs. vancomycin for the treatment of recurrent Clostridium difficile infection. Aliment Pharmacol Ther 2015;41:835–43.
100. Khoruts A, Sadowsky MJ. Understanding the mechanisms of faecal microbiota transplantation. Nat Rev Gastroenterol Hepatol 2016;13:508–16.
101. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proc Natl Acad Sci USA 2005;102:11070–5.
102. Wu S, Chen Y, Li Z, Li J, Zhao F, Su X. Towards multi-label classification: next step of machine learning for microbiome research. Comput Struct Biotechnol J 2021;19:2742–9.
103. Coenen AR, Hu SK, Luo E, Muratore D, Weitz JS. A primer for microbiome time-series analysis. Front Genet 2020;11:310.
104. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science 2011;334:105–8.
105. Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, et al. Daily sampling reveals personalized diet-microbiome associations in humans. Cell Host Microbe 2019;25:789–802 e5.
106. Zmora N, Zilberman-Schapira G, Suez J, Mor U, Dori-Bachash M, Bashiardes S, et al. Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features. Cell 2018;174:1388–405 e21.
107. Berry SE, Valdes AM, Drew DA, Asnicar F, Mazidi M, Wolf J, et al. Human postprandial responses to food and potential for precision nutrition. Nat Med 2020;26:964–73.
108. Quinn TP, Erb I. Examining microbe-metabolite correlations by linear methods. Nat Methods 2021;18:37–9.
109. Sadowsky MJ, Staley C, Heiner C, Hall R, Kelly CR, Brandt L, et al. Analysis of gut microbiota - an ever changing landscape. Gut Microb 2017;8:268–75.