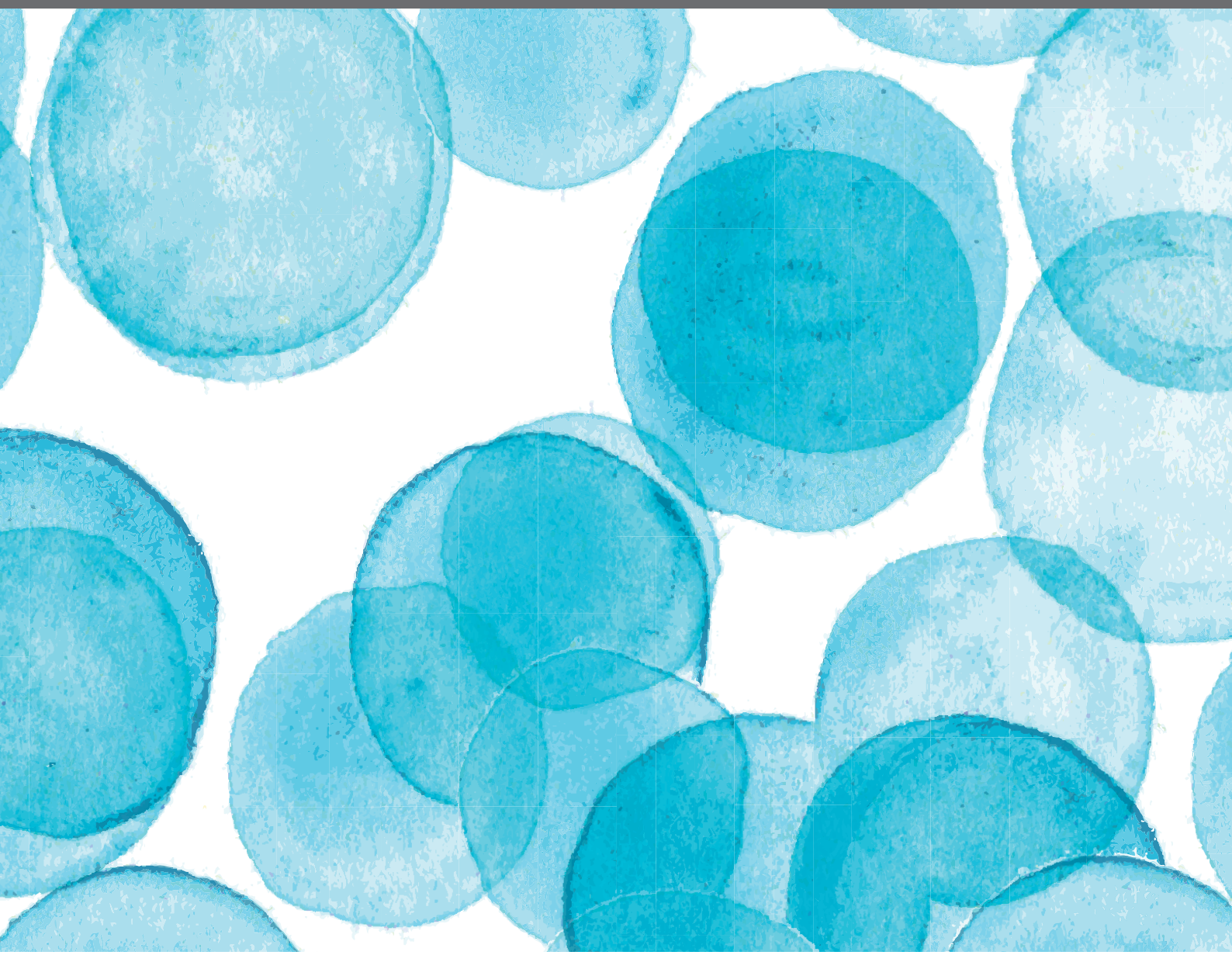
PREDICTIVE MODELING OF HUMAN MICROBIOTA AND THEIR ROLE IN HEALTH AND DISEASE

EDITED BY: Hyun-Seob Song, Steve Lindemann and Dong-Yup Lee
PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-091-8
DOI 10.3389/978-2-88974-091-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

PREDICTIVE MODELING OF HUMAN MICROBIOTA AND THEIR ROLE IN HEALTH AND DISEASE

Topic Editors:
Hyun-Seob Song, University of Nebraska-Lincoln, United States
Steve Lindemann, Purdue University, United States
Dong-Yup Lee, Sungkyunkwan University, South Korea

Citation: Song, H.-S., Lindemann, S., Lee, D.-Y., eds. (2022). Predictive Modeling of Human Microbiota and Their Role in Health and Disease.
Lausanne: Frontiers Media SA. doi:10.3389/978-2-88974-091-8

Table of Contents

frontiers in Microbiology

EDITORIAL
published: 30 November 2021
doi: 10.3389/fmicb.2021.782871

Editorial: Predictive Modeling of Human Microbiota and Their Role in Health and Disease

Hyun-Seob Song[1,2]*, *Stephen R. Lindemann*[3,4] *and Dong-Yup Lee*[5]

[1] Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE, United States, [2] Department of Food Science and Technology, Nebraska Food for Health Center, University of Nebraska-Lincoln, Lincoln, NE, United States, [3] Department of Food Science, Whistler Center for Carbohydrate Research, Purdue University, West Lafayette, IN, United States, [4] Department of Nutrition Science, Purdue University, West Lafayette, IN, United States, [5] School of Chemical Engineering, Sungkyunkwan University, Suwon, South Korea

Keywords: microbial biomarkers, machine learning, microbial association networks, species and gene profiling, precision nutrition

Editorial on the Research Topic

Predictive Modeling of Human Microbiota and Their Role in Health and Disease

OPEN ACCESS

Edited by:
George Tsiamis,
University of Patras, Greece

Reviewed by:
Neil Surana,
Duke University, United States
Samuel C. Forster,
Wellcome Sanger Institute (WT), United Kingdom

***Correspondence:**
Hyun-Seob Song
hsong5@unl.edu

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *24 September 2021*
Accepted: *11 November 2021*
Published: *30 November 2021*

Citation:
Song H-S, Lindemann SR and Lee D-Y (2021) Editorial: Predictive Modeling of Human Microbiota and Their Role in Health and Disease. Front. Microbiol. 12:782871. doi: 10.3389/fmicb.2021.782871

The human microbiota—communities of microorganisms living in or on humans as their host—are deeply involved in various biological processes and functions in our body and play an important role in maintaining physiological and mental health. The criticality of human microbiota is often indicated by the fact that bacterial cells (>100 trillion cells) in our body outnumber human cells (Sender et al., 2016) and carry ∼150 times more genes than the entire human genome (Ursell et al., 2014)—thus called the second genome (Grice and Segre, 2012). Unlike the first genome, which remains largely invariable after being inherited from parents, the second genome's content dynamically changes under a variety of conditions affected by diets, drugs, stress, injuries, and myriad lifestyle and environmental factors. Hence, understanding of how microbiota respond to those perturbations and how the outcomes influence human health is important for developing microbiome-mediated strategies to improve it.

Mathematical modeling of human microbiota and their interactions with host cells under various environmental conditions is indispensable in this regard. The effectiveness of mathematical models has been demonstrated through various applications, including inferring microbe-microbe, microbe-host, and microbe-host-diet interactions (Song et al., 2014; Li et al., 2019; Chowdhury and Fong, 2020), creating new insights into the role of human microbiota in health and disease state (Kumar et al., 2019), and proposing new engineering strategies for intervention (Sheth et al., 2016; Kessell et al., 2020). In addition to process-based models, data-driven modeling is also being popularly used (Marcos-Zambrano et al., 2021) and the combination of these complementary approaches is expected to significantly increase the scope of prediction (Kessell et al., 2020).

Despite promising progress over the past decade, we still lack a complete understanding of the relationships between microbiota (composition and function) and the host (health and disease) under perturbed conditions as well as in homeostasis, and consequently have a limited capability of predicting their interplays. Toward addressing those challenges and promoting new opportunities, this Research Topic collects nine research articles and one review that present the state-of-the-art modeling approaches in various areas of human microbiome research. Below we summarize the contributions from lead computational biologists under the following several categories: discovery of microbial biomarkers and signatures of diseases, quantitative assessment of the impact of bacteria on human health and their association with disease, improved profiling of species and genes in

the microbiome, and incorporation of microbiome data with other clinical features for improved precision nutrition and medicine.

Human microbiota show compositional changes across conditions, including physiological states of the host in health and disease. Particular enrichments of microbes, if observed, can therefore be used as a biomarker of specific diseases. Those signatures could further be considered potential diagnostic and therapeutic targets. Several contributions in the Research Topic address this issue. To predict the risk for developing colorectal metachronous adenoma (MA) after surgical resection, Liu et al. developed a random forest (RF) model using the relative abundance of the gut microbial populations with or without the clinical risk factors. This work is based on the hypothesis that the composition of the gut microbiota before surgery was associated with the risk of developing MA and, therefore, could serve as a potential biomarker for MA. The resulting RF model identified *Escherichia-Shigella* and *Acinetobacter* as key microbiome biomarkers, although the accuracy of prediction was improved when linked with other clinical risk factors such as synchronous adenoma and body mass index. Kort et al. examined how the hypothesis on the use of gut microbiomes as a biomarker can be extended to identify the association of gut microbiota with language development of young children. Using data from rural 3-year-olds in Uganda, they developed regression models by accounting for all possible combinations of three or four species. This comprehensive survey of regression models of all subsets of species led to the identification of *Coprococcus eutactus*, an anaerobic butyrate-producing gut bacterium, as a major predictor of language development in children. In the study of non-small cell lung cancer patients treated with different cycles of osimertinib therapy, Cong et al. identified the shifts in microbial biomarkers between post- and pre-therapy. Through the analysis of intestinal microbial ecological networks constructed by random matrix theory methods, they also found the structure of microbial interaction networks became complicated by including more compact modules in response to osimertinib therapy. A knowledgebase system of the human colorectal cancer (CRC) microbiome constructed by Zhou et al. integrates complementary data and information to improve the predictive power of models in biomarker prediction. The web-based platform allows for systematic inquiry and comparison across different models or databases to identify microbial biomarkers through statistical analysis. The important goal of this platform is to facilitate diagnosis of CRC, identify key factors for clinical transformation, and contribute to the development of cost-effective screening strategies.

Beyond discovering microbiome biomarkers and signatures of disease, another central research challenge is to decipher direct linkages between microorganisms and specific disease types, e.g., through a data-driven network analysis or mechanistic modeling. Along this line, Lei and Wang proposed a new method that enables integrating two similarity-based networks of microorganisms and diseases through the known microbe-disease associations. The resulting integrative network of microbes and diseases may potentially create a new mechanistic understanding of microbe-disease associations that are previously unknown. Compared to existing approaches, the proposed method showed effectiveness in predicting microbe-disease association, as demonstrated through case studies of asthma, chronic obstructive pulmonary disease, and inflammatory bowel disease. A more mechanistic prediction of the bacterial impact on risk factors that may cause disease was made by Bourgin et al. In order to evaluate the impact of the microbial activity vs. host on the cholesterol cycle, they developed a whole-body human model of cholesterol metabolism by incorporating bacterial conversion of bile salts and cholesterol into the existing models that focus on host metabolism. Comprehensive simulations using the model showed that cholesterol conversion to bile salts is the main flux of cholesterol cycle, indicating that bacterial metabolism likely drives cholesterol regulation.

Further, maintenance of the community structure and function of human microbiomes by regulating the ecological balance of microbial populations is key, as disturbances are linked with negative outcomes on human physiological and psychological health. Maintaining a desirable ecological balance of populations in the human microbiome is important because alterations in their composition and function (i.e., dysbiosis) are linked with detrimental physiological and psychological impacts, and result in a wide array of disease conditions. The dynamic model developed by Dedrick et al. can serve as a useful tool to understand bacterial coexistence and stability. The authors constructed an *in silico* model of nasal microbiota composed of up to 20 isolates to predict how the community composition responds to the variation of pH fluctuations in amplitude or frequency. The simulation results showed no significant impact of temporal pH fluctuations on the species coexistence and composition. The numerical model also suggested cooperative interactions among member species that have low niche overlap as a potential mechanism for the observed robustness of nasal microbiota.

Development of predictive human microbiome models is facilitated by advanced gene sequencing technologies, including amplicon sequencing for bacterial composition profiling and shotgun sequencing for metagenomic analyses. As pointed out by Gwak and Rho, it is challenging to perform accurate taxonomic assignment at species level because the 16S rRNA sequences among species in the same genus are highly homologous or even identical. To improve the resolution, they reannotated inconsistent or mislabeled taxa in three major 16S rRNA databases and determined species-level taxonomy using a *k*-nearest neighbor algorithm and the consensus models constructed for each species. In the case studies using salivary and gut microbiome data, the proposed method successfully identified the variation in bacterial composition across different groups based on improved species-level profiling. Ma argued that metagenomic gene abundance data can be analyzed in a similar fashion to operational taxonomic unit analysis by viewing microbiomes as a community of genes, rather than species. Using Taylor's power law, this work analyzed the impact of obesity, inflammatory bowel disease, and diabetes on the human microbiota, highlighting the importance of a sound understanding of metagenomic heterogeneity for the success

of personalized and precision medicine to treat the human microbiome-associated diseases.

Finally, one of the major goals in the human microbiome research is to fundamentally elucidate the complex interactions among diet, gut microbiome, and human health so that we have an improved capability of monitoring wellness states, treating diseases, designing food products, and administering health interventions. Eetemadi et al. provide a review on this issue by discussing a critical role of predictive computational models, particularly machine learning and artificial intelligence. They shared the state of dietary recommendation systems (RSs) and highlight the transition from population-wide to microbiome-aware RSs to provide users with personalized guidelines. They also discussed the details about three complementary approaches for realizing microbiome-aware RS, including knowledge-based, content-based, and collaborative filtering RSs.

The overarching goal of computational modeling of human microbiomes is to enhance our ability to predict their dynamics, association with human disease, and control points that can be used to shape microbiome composition and functions toward improved human health. Some of these control points may target the organisms themselves, others may shape their environments to accomplish these ends. The work contained in this Research Topic uses multiple techniques to provide increased insight into environment-human microbiome-health connections and to substantially advance the field of predictive human microbiome modeling.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

D-YL received funding from the SMC-SKKU Future Convergence Research Program grant and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C2007192).

ACKNOWLEDGMENTS

We thank all the authors and reviewers for their contributions to this Research Topic.

REFERENCES

Chowdhury, S., and Fong, S. S. (2020). Computational modeling of the human microbiome. *Microorganisms* 8:197. doi: 10.3390/microorganisms8020197

Grice, E. A., and Segre, J. A. (2012). The human microbiome: our second genome. *Annu. Rev. Genomics Hum. Genet.* 13, 151–170. doi: 10.1146/annurev-genom-090711-163814

Kessell, A. K., Mccullough, H. C., Auchtung, J. M., Bernstein, H. C., and Song, H. S. (2020). Predictive interactome modeling for precision microbiome engineering. *Curr. Opin. Chem. Eng.* 30, 77–85. doi: 10.1016/j.coche.2020.08.003

Kumar, M., Ji, B. Y., Zengler, K., and Nielsen, J. (2019). Modelling approaches for studying the microbiome. *Nat. Microbiol.* 4, 1253–1267. doi: 10.1038/s41564-019-0491-9

Li, H., Wang, Y. Q., Jiang, J. W., Zhao, H. C., Feng, X., Zhao, B. H., et al. (2019). A novel human microbe-disease association prediction method based on the bidirectional weighted network. *Front. Microbiol.* 10:676. doi: 10.3389/fmicb.2019.00676

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Turukalo, T. L., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12:634511. doi: 10.3389/fmicb.2021.634511

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.* 14:e1002533. doi: 10.1371/journal.pbio.1002533

Sheth, R. U., Cabral, V., Chen, S. P., and Wang, H. H. (2016). Manipulating bacterial communities by *in situ* microbiome engineering. *Trends Genet.* 32, 189–200. doi: 10.1016/j.tig.2016.01.005

Song, H.-S., Cannon, W. R., Beliaev, A. S., and Konopka, A. (2014). Mathematical modeling of microbial community dynamics: a methodological review. *Processes* 2, 711–752. doi: 10.3390/pr2040711

Ursell, L. K., Haiser, H. J., Van Treuren, W., Garg, N., Reddivari, L., Vanamala, J., et al. (2014). The intestinal metabolome: an intersection between microbiota and host. *Gastroenterology* 146, 1470–1476. doi: 10.1053/j.gastro.2014.03.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Microbiology

REVIEW
published: 03 April 2020
doi: 10.3389/fmicb.2020.00393

The Computational Diet: A Review of Computational Methods Across Diet, Microbiome, and Health

*Ameen Eetemadi[1,2], Navneet Rai[2], Beatriz Merchel Piovesan Pereira[2,3], Minseung Kim[1,2,4], Harold Schmitz[5] and Ilias Tagkopoulos[1,2,4]**

[1] Department of Computer Science, University of California, Davis, Davis, CA, United States, [2] Genome Center, University of California, Davis, Davis, CA, United States, [3] Department of Microbiology, University of California, Davis, Davis, CA, United States, [4] Process Integration and Predictive Analytics (PIPA LLC), Davis, CA, United States, [5] Graduate School of Management, University of California, Davis, Davis, CA, United States

Food and human health are inextricably linked. As such, revolutionary impacts on health have been derived from advances in the production and distribution of food relating to food safety and fortification with micronutrients. During the past two decades, it has become apparent that the human microbiome has the potential to modulate health, including in ways that may be related to diet and the composition of specific foods. Despite the excitement and potential surrounding this area, the complexity of the gut microbiome, the chemical composition of food, and their interplay *in situ* remains a daunting task to fully understand. However, recent advances in high-throughput sequencing, metabolomics profiling, compositional analysis of food, and the emergence of electronic health records provide new sources of data that can contribute to addressing this challenge. Computational science will play an essential role in this effort as it will provide the foundation to integrate these data layers and derive insights capable of revealing and understanding the complex interactions between diet, gut microbiome, and health. Here, we review the current knowledge on diet-health-gut microbiota, relevant data sources, bioinformatics tools, machine learning capabilities, as well as the intellectual property and legislative regulatory landscape. We provide guidance on employing machine learning and data analytics, identify gaps in current methods, and describe new scenarios to be unlocked in the next few years in the context of current knowledge.

Keywords: microbiota, gut microbiome, machine learning, artificial intelligence, data analytics, nutrition

OPEN ACCESS

Edited by:
Hyun-Seob Song, University of Nebraska–Lincoln, United States

Reviewed by:
Babak Momeni, Boston College, United States
Matthew John Wade, Newcastle University, United Kingdom

****Correspondence:***
Ilias Tagkopoulos
iliast@ucdavis.edu

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *23 October 2019*
Accepted: *26 February 2020*
Published: *03 April 2020*

Citation:
Eetemadi A, Rai N, Pereira BMP, Kim M, Schmitz H and Tagkopoulos I (2020) The Computational Diet: A Review of Computational Methods Across Diet, Microbiome, and Health. Front. Microbiol. 11:393. doi: 10.3389/fmicb.2020.00393

INTRODUCTION

During the past two decades, the human microbiome has emerged as a biological system with the potential to significantly influence health and disease (Shreiner et al., 2015). Despite our limited understanding regarding its intricate relationship with the host and its environment (Foster et al., 2017), recent discoveries related to the human microbiome have opened new horizons in food science (Barratt et al., 2017), precision medicine (Wishart, 2016), and biotechnology (Taroncher-Oldenburg et al., 2018) among other fields. In parallel, advances in genomics and bioinformatics have provided inexpensive tools to acquire biological and clinical data, as well as the tools to

translate the data into knowledge (Shoaie et al., 2015; Zeevi et al., 2015; Thaiss et al., 2016a; Korem et al., 2017; Baldini et al., 2018; Bauer and Thiele, 2018; Gilbert et al., 2018; Greenhalgh et al., 2018; Knight et al., 2018). Given these advances, the integration of diet, gut microbiome, and human health (DGMH) data has the potential to drive a paradigm shift in the way wellness states are measured, diseases are treated, products are designed, and health interventions are administered. To realize this potential, advances in knowledge are required in order to optimize the composition and metabolic dynamics of microbial communities in relation to desired health and performance outcomes—from dietary interventions and bioengineered products to lifestyle changes and the living environment (**Figure 1**).

In this article, we summarize the research that has been done related to DGMH, with a focus on DGMH data and computational methods. We begin with a brief overview of key areas of current knowledge regarding the interaction between diet, health, and the gut microbiome. We then proceed to review the available data sources and the computational methods currently used, investigate the role that machine learning and artificial intelligence (AI) can play in this area, and summarize the intellectual property (IP) and legislative regulatory landscape. We conclude with recommendations to accelerate research and development efforts through better integration of research resources and tools, especially in the context of computational science and data analytics. A glossary of terms is provided in **Table 6**.

In general, the most recent articles reviewing the computational tools for microbiome data focusing on metagenomic data processing methods provide limited guidance on employing machine learning and data analytics and do not furnish recommendations in the context of DGMH data. The purpose of this manuscript is to help fill this gap by considering relevant literature, describing key challenges and potential solutions, and proposing a framework to improve the potential for research initiatives to accelerate progress in this exciting and potentially revolutionary field.

Current Knowledge: Gut Microbiota and Human Health

Emerging evidence suggests that the intestinal microbiota plays a significant role in modulating human health and behavior [see comprehensive reviews (Sherwin et al., 2018; Pereira et al., 2019; Zmora et al., 2019)]. Several studies have demonstrated that the human intestinal microbiota is seeded before birth (Stinson et al., 2019), and the mode of delivery influences the composition of the gut microbiota (Ferretti et al., 2018; Shao et al., 2019). The gut of a vaginally born newborn is enriched primarily with the vaginal microbiota from the mother, while a cesarean procedure results in the newborn's gut microbiota being dominated by the microbiota of the mother's skin as well as points of contact at the hospital (Dominguez-Bello et al., 2010). Microbial diversity is very dynamic during infancy and increases and converges to an adult-type microbiota by 3–5 years of age (Rodríguez et al., 2015). Evidence is also building to suggest that diet plays a key role in shaping the composition of microbial communities in the infant's gut. For example, species of beneficial bacteria such as *Lactobacillus* and *Bifidobacterium* have been found to be dominant in breastfed infants while species of harmful bacteria such as *Clostridium*, *Granulicatella*, *Citrobacter*, *Enterobacter*, and *Bilophila* have been found to be dominant in formula-fed infants (Bäckhed et al., 2015). In addition, breastfed babies have higher gut microbial diversity compared to formula-fed babies, and several studies indicate that the diversity of bacteria is directly connected to health (Wang et al., 2008; Bäckhed et al., 2015). An unbalanced composition of the infant's gut microbiota has been linked to several childhood diseases, including atopic dermatitis (AD) (Abrahamsson et al., 2012; Zheng et al., 2016) obesity (Yuan et al., 2016), and asthma (Thavagnanam et al., 2008).

The composition of the gut microbiota of an adult human is relatively stable (Shreiner et al., 2015), but several factors can influence it, including antibiotic treatment, long-term change in diet, microbial infections, and lifestyle (Willing et al., 2011; Conlon and Bird, 2015; Mathew et al., 2019; Zmora et al., 2019). Several health conditions are linked to changes in a stable and established gut microbiota such as Crohn's disease (Manichanh et al., 2006), psoriatic arthritis (Scher et al., 2015), type 1 diabetes (de Goffau et al., 2013), atopic eczema (Wang et al., 2008), celiac disease (Schippa et al., 2010), obesity (Castaner et al., 2018), type 2 diabetes (Qin et al., 2012), and arterial stiffness (Menni et al., 2018). However, further research is required to establish direct links between these conditions and the composition of microbial communities in the gut. Interventions, such as oral administration of probiotics/prebiotics and fecal transplants, have shown efficacy on reducing the severity of some conditions, such as diarrhea, acute upper respiratory tract infections, eczema, Crohn's disease, and ulcerative colitis (Anderson et al., 2012; Mansfield et al., 2014; Hao et al., 2015; Saez-Lara et al., 2015; Goldenberg et al., 2017; Delzenne et al., 2019). See **Figure 2** for illustration of factors affecting the gut microbiota.

Data

The increase in size and heterogeneity of information gathered by microbiome studies present great opportunities and serious data analysis challenges (Wooley et al., 2010), with many tools developed to address them (Breitwieser et al., 2017; Quince et al., 2017). These bioinformatics tools quantify low dimensional biological variables, such as the relative abundance of microbial species and metabolites, by using high dimensional data such as DNA sequence reads and mass spectrometry (MS) signatures as illustrated in **Figure 3**. Depending on data quality, sample size, and research hypothesis, different information dimensionalities are used, such as gene-level (Vatanen et al., 2018) or functional gene ontology terms (Brown et al., 2011).

Gut Microbiota Data

Functional characteristics of microbial communities can be revealed using high-throughput metametabolomics (Walker et al., 2014) and metaproteomics (Verberkmoes et al., 2009; Zhang et al., 2018) using MS technologies. Metagenomic and metatranscriptomic content of gut microbiota (which give rise to the functional characteristics) can be quantified using DNA sequencing. The most widely used approach for gut microbiota

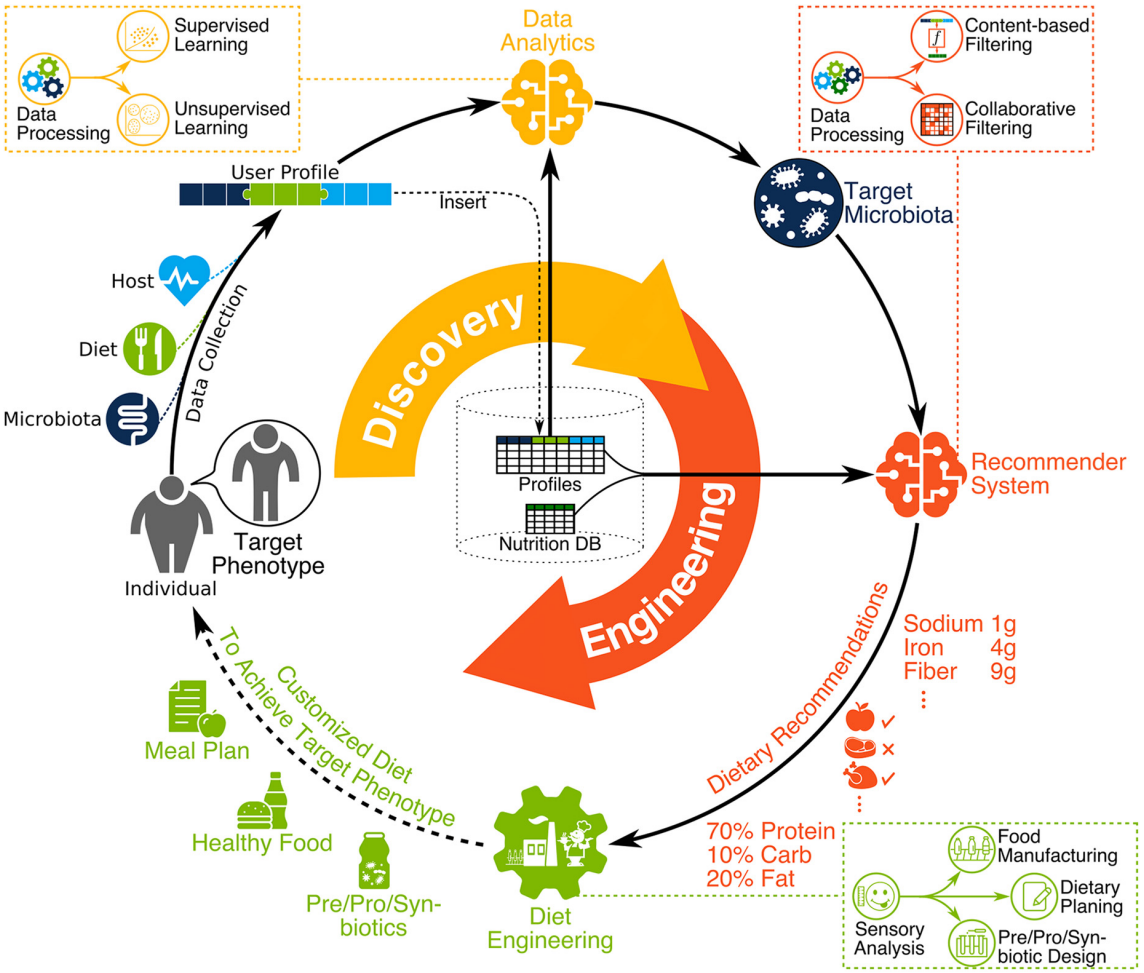


FIGURE 1 | The vision for the next nutrition revolution involves microbiome-aware dietary planning and manufacturing. First, DGMH data is collected, homogenized, and stored, with any new user data integrated as part of a cohesive compendium. Then, DGMH data are analyzed (*data analytics*) to identify the functional characteristics and target microbiota, personalized to the individual and the desired phenotype. This includes data processing followed by supervised and unsupervised learning using a user profile compendium. Bioinformatic tools are used during data processing to extract meaningful information from raw high-throughput data such as metagenomic sequence reads. Then, the *recommendation system* provides dietary recommendations to help achieve target microbiota. This includes the integration of user profiles in a compendium along with nutrition DB proceeded by data processing then content-based and collaborative filtering. Finally, *diet engineering* is performed to create dietary products for the user. This includes the design of prebiotics, probiotics, synbiotics, manufactured food, and detailed dietary planning. In practice, taste and flavor of dietary products is very important to help users commit to any given diet, therefore sensory analysis should inform all dietary engineering efforts.

profiling is *marker gene sequencing*, which relies on sequencing counts of the hypervariable 16S genes to calculate Operational Taxonomic Units (OTUs) (Amann et al., 1995). Searching OTUs against reference databases such as Greengenes (McDonald et al., 2012) and SILVA (Quast et al., 2012) allows inferring relative taxa abundances in a microbiome sample (Langille et al., 2013). *Whole-genome or shotgun metagenomics* (Quince et al., 2017) is a recent technique that not only reveals the microbial community structure, but it can also quantify relative abundances of genes, taxa, conserved functional groups, or over-represented pathways. Within-sample (alpha) and cross-sample (beta) diversity of microbiome can be calculated with respect to genetic, taxonomic, functional, or metabolic pathway profiles of samples (Turnbaugh et al., 2009; Martiny et al., 2011; Huttenhower et al., 2012; Lozupone et al., 2012; Heintz-Buschart and Wilmes, 2017; Ranjan et al., 2018). The Shannon index, Chao1, and Abundance-based Coverage Estimator (ACE) are used to measure alpha diversity while UniFrac, weighted UniFrac, and Bray–Curtis calculate beta diversity. In longitudinal studies, the same measures of diversity, or more sophisticated eigenvalue-based analyses, can quantify the microbiota stability across timepoints (Lozupone et al., 2012; Relman, 2012; Coyte et al., 2015; Mehta et al., 2018). Jackknifing and bootstrapping are used to estimate the bias in diversity estimates, particularly when estimating the number of species (i.e., species richness) in samples (Smith and van Belle, 1984). Some of the most significant publicly available microbiome datasets are listed in **Table 1**.

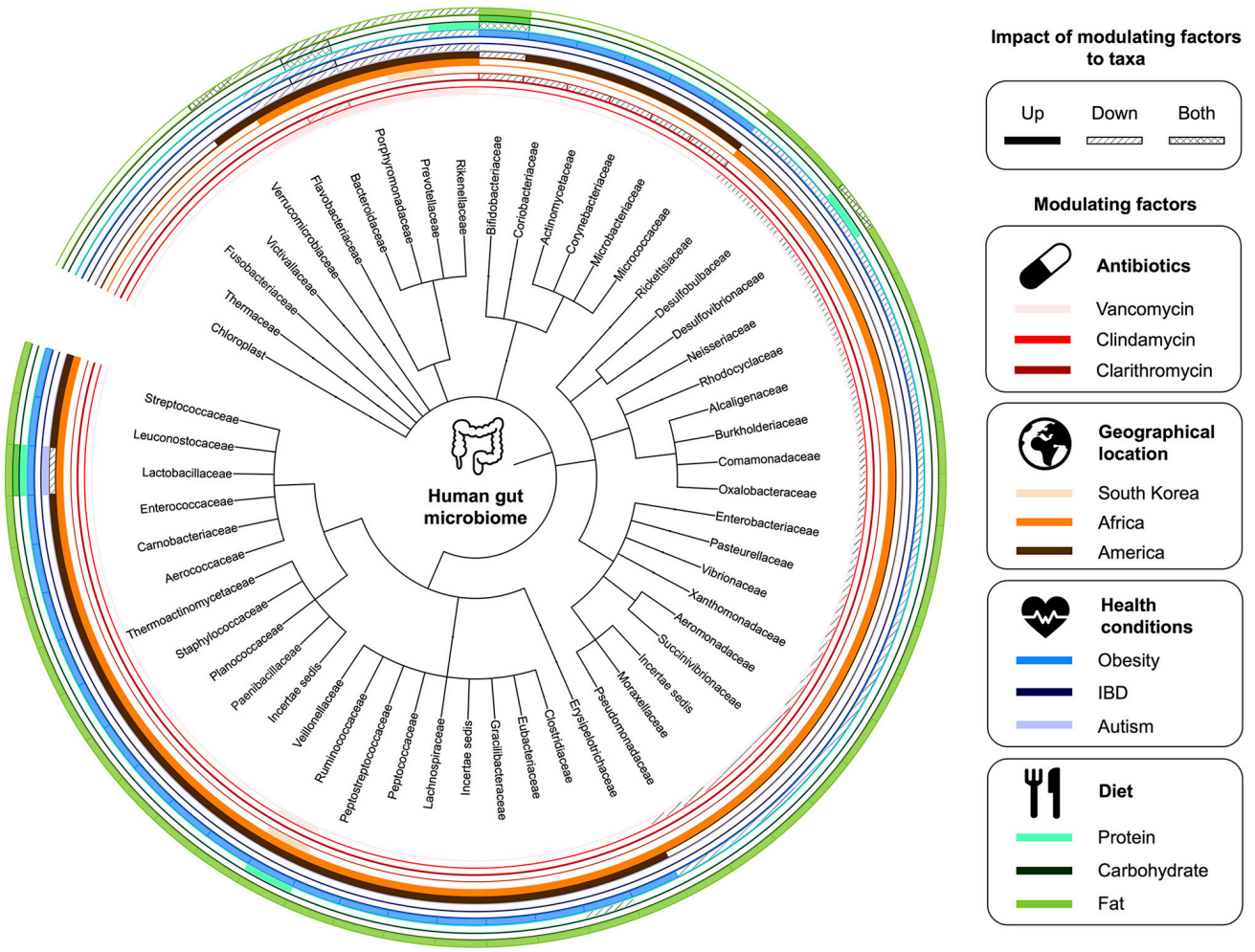


FIGURE 2 | Factors affecting the gut microbiota. A summary of human gut microbiome taxonomy at the family level and the corresponding modulating factors.

Diet Data

Various types of dietary information are collected in gut microbiome studies. This includes fine-grain information such as mass spectrometry (MS) signatures and metagenomic reads (Quinn et al., 2016) or coarse grain information such as dietary style [e.g., Western vs. Mediterranean diet (De Filippis et al., 2016)] from study participants. Diet data collection is often questionnaire-based, either through self-reporting or by a trained interviewer. For self-reporting, a food frequency questionnaire (FFQ) and 24-h dietary recall (24HR) can be used where participants report their dietary intake either every 24 h or over a longer period through a checklist of food items (Shim et al., 2014). A dietary record (DR) can also be used where data collection is done when food is consumed (e.g., using smartphones),which minimalizes reliance on participant's memory. After data collection, the intake amount of macronutrients (fat, carbohydrates, and protein), micronutrients (vitamins and minerals), and food metabolites can be estimated by querying the food items against food composition databases such as the USDA food composition database (US Department of Agriculture and Agricultural Research Service, 2010) and the Canadian nutrient file (Canada, 2010). Note that microbiota of dietary intake can be characterized using metagenomic sequencing as reviewed previously, if not already defined [e.g., probiotics with predefined strains (Sánchez et al., 2017)]. Some studies perform metabolic characterization of dietary intake directly (Quinn et al., 2016), while others rely on pre-characterized metabolic profiles (Zhao et al., 2018). A significant limitation of any analysis is that food composition databases characterize only 0.5% of the known nutritional compounds (Barabási et al., 2019).

Host Data

Profiled host information types can be very high dimensional [e.g., high-throughput genome sequences (Hall et al., 2017)] or low dimensional [e.g., obese vs. non-obese (Thaiss et al., 2014; Cox and Blaser, 2015)]. Host genotype data can come from whole-exome sequencing (WES) (Gopalakrishnan et al., 2018) or a genome-wide association study (GWAS) (Bonder et al., 2016;

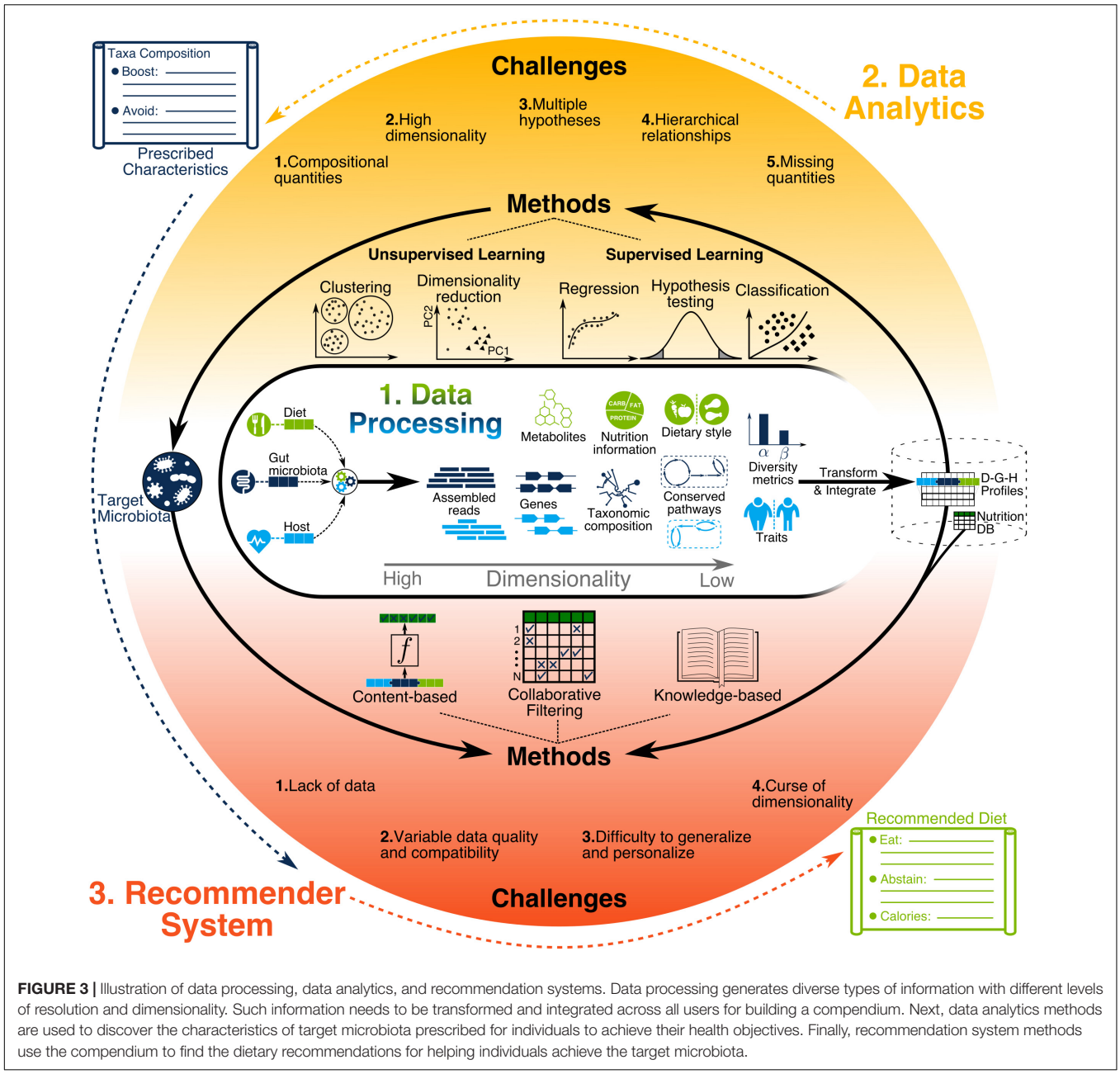


FIGURE 3 | Illustration of data processing, data analytics, and recommendation systems. Data processing generates diverse types of information with different levels of resolution and dimensionality. Such information needs to be transformed and integrated across all users for building a compendium. Next, data analytics methods are used to discover the characteristics of target microbiota prescribed for individuals to achieve their health objectives. Finally, recommendation system methods use the compendium to find the dietary recommendations for helping individuals achieve the target microbiota.

Turpin et al., 2016). It can also be extended by predicting the whole-genome sequence for each individual through genotype imputation software (Howie et al., 2009), as done in several studies (Bonder et al., 2016; Goodrich et al., 2016; Rothschild et al., 2018). Host transcriptomic profiles can be assessed directly using microarrays (Schwartz et al., 2012; de Steenhuijsen Piters et al., 2016) and RNA-Seq (Thaiss et al., 2016b; Pan et al., 2018) or imputed using tools such as PrediXcan (Gamazon et al., 2015) with GWAS data. The genetic and transcriptomic profiles can be summarized into informative lower-dimensional features through gene ontology categories and metabolic pathways using databases such as MetaCyc (Caspi et al., 2017), KEGG (Kanehisa et al., 2011), Reactome (Fabregat et al., 2017), or GO (Antonazzo et al., 2017). Today, limited microbiome studies perform such analysis (Blekhman et al., 2015; Davenport et al., 2015; Dobson et al., 2015). Other important information such as age, gender, ethnicity, body weight, blood pressure, dietary restrictions, and diseases of a host organism can be extracted from medical records, surveys, and interviews.

COMPUTATIONAL ANALYSIS

There have been various reviews concerning microbiome data processing and analysis (Tyler et al., 2014; Tsilimigras and Fodor,

TABLE 1 | Publicly available data from gut microbiota studies.

Project, database, or repository name	Number of cases	Sample types	Disease related (Y/N/B)	Data availability (Y/N/Conditional)	Website
Human Microbiome Project (HMP1)	300	Nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract	N	Y	NIH Human Microbiome Project - Home, 2019
Integrative Human Microbiome Project (iHMP): pregnancy and preterm birth (MOMS-PI)	∼2,000	Mouth, skin, vagina, and rectum	Y	Y	NIH Human Microbiome Project - Home, 2019
Integrative Human Microbiome Project (iHMP): onset of IBD (IBDMDB)	∼90	Stool and blood	Y	Y	NIH Human Microbiome Project - Home, 2019
Integrative Human Microbiome Project (iHMP): onset of type 2 diabetes (T2D)	∼100	Fecal, nasal, blood, serum, and urine	Y	Y	NIH Human Microbiome Project - Home, 2019
American Gut Project (AGP)	>3,000	Stool and swabs from skin/mouth	B	Y	American Gut, 2019
Personal Genome Project microbiota component (PGP)	>5,000	Skin/oral/fecal	–	Y	Data – The Harvard Personal Genome Project (PGP), 2019
TwinsUK	>11,000	Multiple	–	C	TwinsUK, 2019
Global Gut Project (GG)	531	Fecal	N	Y	Yatsunenko et al., 2012
Project CARDIOBIOME	>4,000	–	–	N	
Pediatric Metabolism and Microbiome Repository (PMMR)	∼350	Human microbial cell lines, stool, and/or DNA and RNA	Y	N	https://clinicaltrials.govClinicalTrials.gov, 2019
Lung HIV Microbiome Project (LHMP)	162	Lung, nasal, and/or oropharyngeal cavities	Y	Y	BioLINCC, 2019
The Study of the Impact of Long-Term Space Travel on the Astronauts' Microbiome (Microbiome)	9	Saliva and gastrointestinal	N	N	NASA, 2019
Michigan Microbiome Project (MMP)	–	–	–	N	The Michigan Microbiome Project, 2019
uBiome	–	Gut, mouth, nose, genitals, and skin	B	C	
Human Oral Microbiome Database (eHOMD)	–	Upper digestive and upper respiratory tracts, oral cavity, pharynx, nasal passages, sinuses, and esophagus	–	Y	HOMD : Human Oral Microbiome Database, 2019
Human Pan-Microbe Communities (HPMC)	>1,800	Gastrointestinal	B	Y	HPMCD: Human Pan Microbial Communities Database, 2019
Curated Metagenomic Data	>5,000	Multiple	B	Y	curatedMetagenomicData, 2019
European Nucleotide Archive	–	–	–	Y	European Nucleotide Archive EMBL-EBI, 2019
EBI-metagenomics portal samples	>20,000	Multiple	B	Y	EMBL-EBI Mg, 2019
MG-RAST	>10,000	Multiple	B	Y	MG-RAST, 2019

2016; Breitwieser et al., 2017; Quince et al., 2017; Knight et al., 2018). Here we focus on data analytics, machine learning, and AI-based recommendation system methods that enable microbiome-aware systems involving diet and wellness. We provide readers insight into important methods, challenges that arise, suggested solutions as well as blueprints of example scenarios to be used in their research. See Qu et al. (2019), Topçuoglu et al. (2019), and Zhou and Gallins (2019) for further explanation and examples of the machine learning methods discussed here.

Microbiome Data Processing Tools

There are a substantial number of publicly available microbiome data processing methods and pipelines that can generate the various types of data discussed. **Table 2** provides a representative summary of such methods and pipelines. For 16S data, QIIME (Caporaso et al., 2010) and MOTHUR (Schloss et al., 2009) provide a wider range of options for the user compared to UPARSE (Edgar, 2013), but all are popular pipelines. QIIME 2 (Bolyen et al., 2019) is now emerging as a powerful replacement

TABLE 2 | A summary of highlighted methods and pipelines for microbiome data processing.

Steps	Sub-step descriptions	Highlighted methods and their availability in popular pipelines (**Q**IIME, **M**OTHUR, and **U**PARSE)
(1) Quality control	Chimera removal and noise mitigation	Trimmomatic$^{(Q)}$ (Bolger et al., 2014), AmpliconNoise$^{(Q,M)}$ (Bragg et al., 2012), UNOISE$^{(M,U)}$ (Edgar, 2016), UCHIME$^{(Q,M,U)}$ (Edgar et al., 2011), Deblur$^{(Q,M)}$ (Amir et al., 2017), and DADA2$^{(Q)}$ (Callahan et al., 2016)
	Remove host DNA contaminant reads	Bowtie2$^{(Q)}$ (Langmead and Salzberg, 2012), BMTagger (Agarwala and Morgulis, 2011), and DeconSeq (Schmieder and Edwards, 2011)
(2) Sequence assembly	*De novo* read assembly	MEGAHIT (Li et al., 2015), MAFFT$^{(Q,M)}$ (Katoh and Standley, 2013), UCLUST$^{(Q,U)}$ (Edgar, 2010), and metaSPAdes$^{(Q,M)}$ (Nurk et al., 2017)
	Read alignment to annotated database	DIAMOND (Buchfink et al., 2014), NAST$^{(Q,M)}$ (DeSantis et al., 2006), USEARCH$^{(Q,U)}$ (Edgar, 2010), and VSEARCH$^{(Q,M)}$ (Rognes et al., 2016)
(3) OTU analysis	Assignment of reads to OTUs	UPARSE-OTU$^{(U)}$ (Edgar, 2013), Kraken (Wood and Salzberg, 2014), MetaPhlAn2$^{(Q)}$ (Truong et al., 2015), and DOTUR$^{(M)}$ (Schloss and Handelsman, 2005)
(4) Functional profiling	Functional profiling and prediction	MEGAN (Huson et al., 2016), HUMAnN (Abubucker et al., 2012), MetaCLADE, MOCAT (Kultima et al., 2016), and PICRUSt (Langille et al., 2013)
(5) Diversity analysis	Diversity, evenness, and richness metrics	Alpha [e.g., Chao1$^{(Q,M,U)}$] and Beta [e.g., Jaccard$^{(Q,M,U)}$]

to its predecessors, partly due to its extensibility and support. For whole metagenomic sequencing, methods such as Kraken (Wood and Salzberg, 2014), MEGAN (Huson et al., 2016), MetaPhlAn2 (Truong et al., 2015), and HUMAnN (Abubucker et al., 2012) are used.

Challenges in Microbiome Data Processing

Growth in the variety and complexity of data processing tools presents opportunities but also significant challenges for new investigators. First, although best practices have been suggested (Knight et al., 2018), tools are still far from a fully automated user experience that would lead to reliable results. Second, microbial genomes with different abundances are sequenced together, making metagenomic assembly more challenging compared to single genome assembly where the sequence coverage is approximately uniform. Third, the number of uncharacterized microbes (known as microbial dark matter) exacerbates problems associated with unaligned and misaligned sequence reads. Fourth, evaluation of methodology and findings from different studies is difficult since each study may use a different method or a different implementation of the same method in their data processing pipeline. Fifth, data collection and integration of microbiome data from different studies are difficult because of many factors including differences in wet-lab library preparation (e.g., primers used), differences in sequencing devices and their settings (e.g., coverage), and non-uniform methods of formatting and storage for microbiome data and metadata. See Quince et al. (2017) for further discussion concerning microbiome data processing challenges.

Data Analytics and Machine Learning

Data processing is considered to be the step necessary for converting the raw data, such as metagenomics sequence reads, into biologically meaningful representations, such as OTU counts using bioinformatics tools, some of which are done in the sequencing device itself. Data analytics, start after the integration of processed sample data from various information sources (i.e., microbiota, diet, and host), as illustrated in **Figure 3**. In most cases, all samples are from a single study, which helps ensure consistency with respect to the experimental settings and data processing protocols used. Furthermore, limited resources force the researchers to narrow their data collection to particular information types in order to have sufficient statistical power for hypothesis testing. A recent increase in the number of microbiome studies with publicly available data has enabled cross-study data integration (Pasolli et al., 2016, 2017; Duvallet et al., 2017; Wang et al., 2018; Thomas et al., 2019; Wirbel et al., 2019). In such cases, extra precautions are necessary to minimize biases introduced by inconsistencies among datasets during data collection, sample preparation, sequencing, and data processing.

Challenges in Microbiome Data Analysis

A number of challenges arise when analyzing microbiome data, as summarized in **Table 3**. The first challenge is due to *compositional quantities* in microbiome data. Quantities such as the number of reads assigned to a given species, which can only be interpreted as proportions, are called compositional. These quantities cannot be compared directly across multiple samples. Conclusions should not be made based on the number of reads assigned to individual sample features (e.g., OTUs, genes, and functional groups) since they do not represent absolute abundances due to instrumental limitations (Gloor et al., 2017). Instead, the assigned number of reads should be converted to relative abundances and analyzed with that in mind. Some studies perform rarefaction to adjust for differences in library size due to unexhaustive metagenomic sampling. Although several pipelines provide this functionality, it has been found inadmissible for metagenomics microbiome studies as it discards many reads leading to decreased sensitivity in differential abundance testing (McMurdie and Holmes, 2014) and biased estimates for alpha diversity (Willis, 2019). The second challenge is due to the *high dimensionality* associated with OMICS data. Datasets in which items are characterized by a high number of features while the number of items is limited are called high dimensional. In microbiome studies, a limited number of individuals are characterized using many host, diet, and microbiome features leading to high dimensional datasets (Li, 2015). Dimensionality can be reduced by grouping OTUs into phylogenetic variables, regularization, or unsupervised dimensionality reduction (DR) (explained below). The third challenge is about testing *multiple hypotheses* in an exploratory analysis. It relates to the fact that, as the number of hypotheses

TABLE 3 | Key challenges that arise in microbiome data analysis with examples and suggested solutions.

Challenges in microbiome data analysis	Examples and solutions
(1) Compositional quantities: Metagenomic data processing provides read counts for discovered entities such as genes, species, and OTUs from a given sample. These read counts are only meaningful within a sample.	**Example:** Metagenomic analysis of feces samples tells us that Person A has 5 reads mapped to bacterium *Escherichia coli,* while person B has 10. Can we conclude that this bacterium is more populated in the gut of person B compared to person A? *Answer*: No, read counts cannot be compared across samples. **Solutions:** (I) Convert read counts to relative abundances before comparison. (II) If an optimization problem is defined using read counts, add constraint for total counts per sample.
(2) High dimensionality: Metagenomic data processing results in many entities such as genes and species discovered for each sample, which may not be shared among multiple samples. During data aggregation, one dimension is associated to each entity resulting in a high number of dimensions compared to the number of samples.	**Example:** Metagenomic data processing of feces samples from 20 individuals results in relative abundances for 10 microbial families per sample. Can we use classical linear regression to predict an individual's age using relative abundances from aggregated data? *Answer:* No, aggregating 20 samples results in more than 20 microbial families. **Solutions:** (I) Use dimensionality reduction such as PCA prior to regression. (II) Use regularized linear regression such as Lasso. (III) Use microbial abundances of higher-order taxonomic ranks such as phylum instead of family.
(3) Multiple hypotheses: The high-dimensional nature of metagenomic data allows the researcher to generate a large number of hypotheses, which leads to seeing patterns that simply occur due to random chance. This is sometimes called "the high probability of low probability events."	**Example:** Metagenomic data processing provides relative microbial abundances at species level using feces samples of 200 individuals, half of which are diagnosed with Crohn's disease and the rest are healthy. Performing a *t*-test identifies that the relative abundance of 40 species (amongst 1,000) are significantly different between microbiota of sick and healthy individuals ($p\text{-value} < 0.05$). Is this result correct? *Answer:* No, the standard threshold of 0.05 for *p-value* is only acceptable when a single hypothesis is involved while the *t*-test is performed 1,000 times leading to many false discoveries. **Solution:** Calculate FDR adjusted *p-value* (i.e., *q-value*) of 0.05 to control the false discovery rate.
(4) Hierarchical relationships: Assumptions of independence do not hold in microbiome data since taxonomic variables (e.g., species and OTUs) have known hierarchical relationships due to genetic and phenotypic similarities. Therefore, common statistical techniques that assume independence between variables are problematic.	**Example:** Beta-diversity can be used to calculate the similarity between groups of microbiome samples. Can we simply calculate the Beta-diversity using standard Euclidean distance between relative abundances at a given taxonomic order? *Answer:* No, Euclidean distance doesn't take into account the similarity between species. **Solution:** Use phylogeny-aware metrics such as UniFrac distance instead, which takes into account the phylogenetic tree when calculating distances.
(5) Missing quantities: Metagenomic data often lacks information about the functions of the microbial communities which can only be estimated using meta-transcriptomics or meta-proteomics. However, deciphering microbiota's function is a major goal in microbiome studies.	**Example:** In one case, metagenomic data processing from marker-gene data has provided us with relative abundances at the genus level, but we do not know the possible functions of the microbiota in terms of proteins that it can produce. Should we abandon further analysis? *Answer:* No, although we don't have direct information about proteins, we can infer. **Solution:** Databases such as Greengenes contain the whole-genome sequence of identified species at various taxonomic orders which can be used for gene and protein inference.

increases, the chance of false discoveries also increases. This can be addressed by increasing sample size and *p*-value adjustment (explained below). The fourth challenge relates to *hierarchical relationships* amongst bacterial species due to their shared ancestors. Assumptions such as independence among samples may not hold, leading to wrong estimations of correlation (Felsenstein, 1985) and phylogeny-aware methods to address the issue. The fifth challenge is about *missing quantities* in sampled data. For example, when marker gene sequencing is used, quantities relating to the amounts of functional genes in the microbiome are not directly available (i.e., missing). Identifying functions of microbial organisms is important for understanding the gut microbiota. Such information can be estimated using metatranscriptomics data, which is often not available. Data imputation tools, such as PICRUSt (Langille et al., 2013), help to mitigate this through gene imputation based on annotated databases.

The methods for identifying microbiota characteristics associated with host phenotypes of interest can be categorized into two main groups, based on whether they use supervised or unsupervised learning. Supervised learning methods require labeled data, while unsupervised learning methods can be used when records are not labeled. More advanced methods include semi-supervised learning (Zhu, 2005), which takes advantage of both labeled and unlabeled data, and transfer learning (Pan and Yang, 2010), which transfers knowledge learned from one task to another, are not discussed here.

Supervised Learning Methods

Hypothesis testing and variation analysis

Analysis of variation may involve single or multiple variables. For a single variable hypothesis, the student's *t*-test or non-parametric tests, such as Wilcoxon rank-sum or Kruskal–Wallis, can be used. For example, the *t*-test has been used to show that patients with ADHD have a lower alpha-diversity index of gut microbiota compared to healthy controls (Prehn-Kristensen et al., 2018). Non-parametric tests are good alternatives when the assumptions regarding the data being normally distributed do not hold. For example, the Wilcoxon rank-sum test is used on predicted pathway data, suggesting that enzymes in the "Glycan Biosynthesis and Degradation" pathway increase in summer when compared to winter (Davenport et al., 2014). In cases where a statistical test is repeated with different hypotheses (i.e., multiple hypothesis testing), the statistical significance should be adjusted by methods such as an FDR adjustment (i.e.,

q-value) (Benjamini and Hochberg, 1995) or Holm's procedure (Rice, 1989).

When the hypothesis that is investigated contains multiple variables, MANOVA (Smith et al., 1962) or non-parametric alternatives such as PERMANOVA (Anderson, 2001) or ANOSIM (Clarke, 1993) can be used. The samples are first assigned to multiple groups (e.g., based on some feature values). The goal is to quantify how much this grouping can explain the distribution of values in any given sample feature (response variable). The simplest case is the popular method called analysis of variance (ANOVA), which considers a single response variable with a normal distribution. For instance, in a recent study, two bacterial phyla (Bacteroidetes and Firmicutes) were identified using ANOVA with different relative abundance in the microbiota of children living in a rural African village compared to European children (De Filippo et al., 2010). ANOVA can be generalized to multivariate analysis of variance (MANOVA), which can have multiple response variables. For example, it is used to investigate the overall difference in composition between the microbiota of children with Prader–Willi syndrome and children with simple obesity, before and after treatment (Zhang et al., 2015). In many cases, normal distribution assumptions do not hold; hence, non-parametric methods are used. In one study, PERMANOVA is used to detect taxonomic differences in the microbiota of patients with Crohn's disease when compared to healthy controls (Pascal et al., 2017).

Regression and correlation analysis

A general understanding of the extent of association among pairs of variables can be achieved using correlation analysis. Correlation metrics measure different types of relationships. For example, the Bray–Curtis measures abundance similarities (Bray and Curtis, 1957), the Pearson correlation coefficient quantifies linear relationships, and the Spearman correlation coefficient quantifies rank relationships (Spearman, 1904). In (Weiss et al., 2016), the authors perform a simulation-based comparison on a range of correlation metrics for microbiome data. Metrics such as SparCC (Friedman and Alm, 2012) and LSA (Ruan et al., 2006) perform particularly better as they are designed to capture complex relationships in compositional microbiome data. For example, SparCC is used for analyzing the TwinUK dataset to identify bacterial taxa whose abundances are influenced by host genetics (Goodrich et al., 2014). This was done by creating a correlation network between microbial families based on their intraclass correlation. More recently, the phylogenetic isometric log-ratio (PhILR) transform has been introduced (Silverman et al., 2017) to transform compositional data into non-compositional space where standard data analytic techniques are applicable. Usage of such transformations should be limited to features that are compositional and phylogenetic in nature.

Regression methods aim to predict the change in one continuous variable using other variables. Correlation analysis can be considered a special case of regression with a single input variable. Standard linear regression can be used for various DGMH predictive tasks. However, when variables relate to OTU abundances, the typical assumptions of a linear relationship, normal distribution, and dependence do not hold. For example, when the goal is to predict the composition of OTUs [normalized for summing up to one (Tyler et al., 2014)], zero-inflated continuous distributions are used. Often a two-part regression model is used where part I is a logistical model to calculate the probability that the given OTU is present. Part II is a generalized linear regression using beta distribution to predict relative abundance assuming the presence of OTU in the sample (Ospina and Ferrari, 2012; Chen and Li, 2016; Peng et al., 2016). Phylogenetic comparative methods (PCMs) such as phylogenetic generalized least squares (PGLS) are used to control for dependence among observations given the phylogenetic hierarchies (Washburne et al., 2018). Ignoring the phylogenetic ancestry of microbial species can increase the chance of false discovery in regression models (Felsenstein, 1985). PCMs are not widely used in microbiome studies today, which may be one reason for a high number of false positives that can be alleviated by using them (Bradley et al., 2018).

Canonical correlation analysis (CCA) can be used (Hotelling, 1992) to investigate the correlation between two groups of variables (e.g., abundances of microbiome OTUs and metabolites). CCA finds linear transformation pairs that are maximally correlated when applied to data while ensuring orthogonality for different transformation pairs. The original CCA, however, fails for high dimensional microbiome data when the number of variables exceeds the number of samples. This can be addressed using regularization, giving rise to sparse CCA methods (Witten et al., 2009). For example, a sparse CCA is applied to investigate correlations between the gut microbiome and metabolome features in type 1 diabetes (Kostic et al., 2015).

Classification

In supervised classification, the goal is to build a predictive model (classifier) using labeled training data. The labels can have binary or categorical values (in contrast to regression where labels are continuous and numerical). In one study, a classifier was built to predict the geographical origin of sample donors using relative OTU abundances estimated from 16s rRNA gut samples (Yatsunenko et al., 2012). This was done using the method called Random Forests (RF), which constructs an ensemble of decision trees (Breiman, 2001). In a different study, the classification task was to identify healthy vs. unhealthy donors given relative OTU abundance data (including species level) coming from shotgun metagenomics sequencing of the gut (as well as other body sites) (Pasolli et al., 2016). In addition to RF, they used the support vector machine (SVM) classifier, which is a powerful method for building generalizable and interpretable models and is mathematically well understood (Suykens and Vandewalle, 1999). In their study, RF classifiers performed better than SVM except in a few datasets. Both RF and SVM have built-in capability to deal with overfitting issues that arise in high-dimensional datasets. RF achieves this using an ensemble-based technique where the prediction is made based on predictions from many trained classifiers. In SVM, parameters of the predictive model are constrained based on *a priori* defined criteria. Note that constraining the model parameters is often mathematically equivalent to regularization (Scholkopf and Smola, 2001). In both cases, the objective is to minimize the

value of a loss function that calculates the overall error in model predictions. When regularization is used, the loss function not only depends on prediction errors but also on the magnitude of model parameters. For example, in L1 regularization, the absolute values of model parameters are scaled and added to the loss function. Therefore, when two models have a similar error, the model with smaller parameter values will be selected during training. L1 regularization is commonly used for feature selection by picking only the non-zero features of the trained model because such a model achieves a low prediction error while using a subset of features.

Artificial neural networks (ANN) can also be used for classification and are shown to outperform other techniques in many areas of biology (Kim et al., 2016, 2017; Singh et al., 2016; Eetemadi and Tagkopoulos, 2018) as well as computer vision and natural language processing, to name a few (LeCun et al., 2015). Recently, a new ANN-based method called Regularization of Learning Networks (RLN) was designed and evaluated microbiome data. RLN provides an efficient way for tuning regularization parameters of a neural network when a different regularization coefficient is assigned for each parameter (Shavitt and Segal, 2018). They use RLN to predict human traits (e.g., BMI, cholesterol) from estimated relative OTU abundances and gene abundances. We expect the development of new classification methods that can deal with the aforementioned challenges arising in DGMH data by considering the biological phenomenon, properties of measurement instruments, and upstream data processing pipelines.

Unsupervised Learning Methods

Dimensionality reduction

High-dimensional datasets can provide a high resolution and multifaceted view of a phenomenon such as gut microbiota. Predictive performance in data analytics can increase significantly given such data. Many data analytics methods, however, fall short when presented with high-dimensional data that necessitates using DR. Once dimensionality is reduced, data visualization and analytics become more accessible. Principal component analysis (PCA) is one of the most widely used DR methods. It replaces the original features with a few uncorrelated features called principal components (PCs), which are linear combinations of the original features and preserve most of the variance within the data. In one study, PCA was applied to predicted abundances of about 10 million genes from the gut microbiota of donors (Li et al., 2014). Reducing dimensionality from 10 million to two dimensions only enabled visualization of data on a standard two-dimensional scatter-plot (i.e., PCA plot) showing a clear distinction between the microbiota of Danish and Chinese donors. In another study, the top five PCs of individual bacteria's genome (sequenced from infant fecal samples) were used to create a classifier for predicting antibiotic resistance (Rahman et al., 2018).

The relationships among features in a microbiome study can be used in DR, giving rise to various factor analysis (FA) methods we review here briefly. Multiple factor analysis (MFA) is an extension of PCA that considers predefined grouping of features during DR to ensure equal representation for all groups of features in derived PCs (Abdi et al., 2013). In one study (Robertson et al., 2018), MFA is used for simultaneous 2D visualization of host and microbiome features (see Morgan et al., 2012; Raymond et al., 2016 for other examples). Exploratory factor analysis (EFA) is used to identify unobserved latent features called factors to explain the correlations among observed features (Yong and Pearce, 2013). Factors that are identified by EFA are uncorrelated to each other similar to PCs in PCA; however, PCs are used to explain overall variance instead of correlations. EFA has been used in a recent study to extract four factors explaining the correlations among 25 top taxa for studying the association of microbiome with early childhood neurodevelopmental outcomes in 309 infants (Sordillo et al., 2019). Confirmatory factor analysis (CFA) and structural equation modeling (SEM) can be used to examine the extent to which a hypothesized model of latent features and their relationships with observed variables are supported by the data (Schreiber et al., 2006). In a recent study, a theoretical framework is proposed and examined using CFA to model the influence of maternal and infant factors on the milk microbiota (Moossavi et al., 2019). The R packages lavaan (Rosseel, 2012) and FactoMineR (Lê et al., 2008), as well as the IBM SPSS software (IBM Corp, 2013), are widely used for factor analysis.

Another related method is principal coordinate analysis (PCoA), also called multidimensional scaling (MDS) (Kruskal, 1964), which is commonly employed for 2- and 3-dimensional visualization of beta diversity. It can deal with situations where distances between individual feature vectors from samples cannot be used directly (e.g., due to significant sparsity and phylogenetic relationships). PCoA takes a matrix of distances among samples (e.g., UniFrac distance between OTU abundances of a pair of sample donors) as input. It then assigns new coordinates such as PC1 and PC2 to each sample such that the Euclidean distances in the new coordinate are similar to the ones in the matrix. For example, PCoA was applied given UniFrac distances between OTU abundances (from 16S rRNA samples) from the gut microbiota of donors (Yatsunenko et al., 2012). Two-dimensional visualization using PC1 and PC2 showed that the gut microbiota of donors who lived in the United States is distinct from the gut microbiota in donors living in Amerindian and Malawian villages.

Linear discriminant analysis (LDA) is also a DR technique, although supervised and closely related to regression and ANOVA. Unlike PCA and PCoA, it requires class labels. It generates new features that are linear combinations of the original ones while separating samples with respect to their class labels. In one study, LDA was used to distinguish gut microbiota samples based on diet but not for DR (Paulson et al., 2013). Successful usage of LDA for high dimensional microbiome data may require regularization to account for overfitting as similarly used for high-dimensional microarray (Guo et al., 2006).

The optimal amount of reduction in dimensionality (e.g., the number of principal components) varies given the data and the task downstream. For data visualization tasks, it is largely constrained by the limitations of human visual perception (three dimensional). For downstream supervised learning tasks, we are often interested in the maximum amount of DR without

a significant decrease in predictive power. This is showcased in Bartenhagen et al. (2010), where the impact of the amount of DR on classification performance is evaluated for gene expression data.

Cluster analysis

Similar microbial communities are expected to exhibit analogous effects on the host organism (Gould et al., 2018). Once a similarity measure is defined, various cluster analysis methods can be used to find groups of samples with similar microbiota. In one study, three robust microbiota clusters (called enterotypes) were identified using cluster analysis from 16s rRNA data of fecal samples (Arumugam et al., 2011). It was later shown that such clustering results are not only sensitive to data but also to choices made during analysis (Koren et al., 2013). We enumerate four important choices impacting cluster analysis results (other than upstream data processing). First is the distance measure. Standard distance metrics such as the Euclidean and Manhattan distance are simple, well understood, and supported in many clustering libraries. Applicability of such metrics depends on prior compositionality aware transformations such as ILR. Beta-diversity metrics such as weighted and unweighted UniFrac distances are designed for microbiome analysis considering compositionality and phylogenetic dependencies of microbiome data. Researchers should pay attention to the properties of the distance metric used in order to better understand the clustering results. Second is the clustering algorithm. Algorithms such as Partition Around Medoids (Kaufman and Rousseeuw, 1987) and Hierarchical Clustering (Murtagh and Contreras, 2012) can employ various distance metrics. Others, such as *k*-means, are tied to a single distance measure but computationally less demanding. Third is the number of clusters. Clustering algorithms often require the number of clusters to be provided as input. When unknown, the number that provides higher cluster scoring is picked. Prediction strength (Tibshirani and Walther, 2005), silhouette index (Rousseeuw, 1987), and Calinski–Harabasz (Caliński and Harabasz, 1974) are popular cluster scoring metrics. Fourth is the method used to identify the robustness of clustering results. Often a cluster scoring metric that is not used to identify the number of clusters is used as a robustness measure. Recent studies consider the effect of the above choices during cluster analysis to better understand how results can be generalized (Hildebrand et al., 2013; Costea et al., 2018).

The integration of data from disparate omics data types (also called integrative omics) and other heterogeneous metadata enables a more comprehensive look into the underlying biology (Karczewski and Snyder, 2018). Integrative omics data analysis methods have been categorized into three types (Kim and Tagkopoulos, 2018). First is *data-to-data*, where disparate data types are analyzed together. For example, CCA can be used to investigate the correlations between metagenomics and metabolomics data, as discussed before. Second is *data-to-knowledge*, where the knowledge gained from analyzing some data types are used to inform analysis of other data types. For example, a metagenomics analysis of colon cancer patients can lead to further investigation of candidate genes using targeted proteomics analysis. Third is *knowledge-to-knowledge*, where the data types are initially analyzed separately, but the acquired knowledge is integrated together afterward to either identify hypotheses that are supported by multiple data types or create a more complete view of a given phenomenon. For example, differentially expressed genes and differentially abundant metabolites in the digestive tract of patients with Crohn's disease can be used together for narrowing down pathways involved in disease etiology. See Huang et al. (2017), Karczewski and Snyder (2018), Kim and Tagkopoulos (2018), and Jiang et al. (2019) for comprehensive reviews.

Recommendation Systems and Artificial Intelligence

The human microbiome is referred to as "our second genome" and has a major influence on our health (Grice and Segre, 2012). Although it is known for its resilience (Lozupone et al., 2012; Relman, 2012), unlike the human genome, it has considerable plasticity hence providing ample opportunities in the design of new types of food, medical interventions, and dietary recommendations (Gentile and Weir, 2018). Despite recent progress in microbiome research, switching from population-wide dietary recommendations to microbiome-aware recommendations is not yet realized. See **Table 4**, for a representative summary of recent microbiome-aware diet recommendation studies. Once a personalized healthy target microbiome is identified using data analytics methods, a recommendation system (RS) can utilize this information to suggest the path toward establishing it in the host and ensuring the health benefits. One approach is to use a knowledge-based RS where recommendations are made using a limited number of approved drugs and dietary prescriptions. Although this would be a good starting point, such a system would be limited in its ability to provide precise and personalized recommendations that usually need a platform that can create new products or processes on a case-by-case basis. Recent studies simulate a virtual gut microbiome by integrating known metabolic pathways of microbial species with the individual's microbiome and diet (Shoaie et al., 2015; Baldini et al., 2018; Bauer and Thiele, 2018; Greenhalgh et al., 2018). Such mechanistic modeling is very promising, however, it is currently hindered by numerous challenges, such as incomplete characterization of an individual's gut and metabolic pathways of their microbiome. There is considerable research on AI-based RS related to food, drug design, and health (Tran et al., 2017; Suphavilai et al., 2018), but its application with microbiome data is in its early stages (Zeevi et al., 2015; Thaiss et al., 2016a; Korem et al., 2017). Commercial investments in this area have already started, with companies such as UBiome and DayTwo using 16S rRNA technology to provide insights into our personal microbiota and suggest dietary recommendations.

Recommendation system is defined as "any system that guides a user in a personalized way to interesting or useful objects in a large space of possible options or that produces such objects as output" (Burke, 2002). Microbiome-aware diet recommendations can be generated from knowledge-based, content-based, or collaborative filtering, as described next.

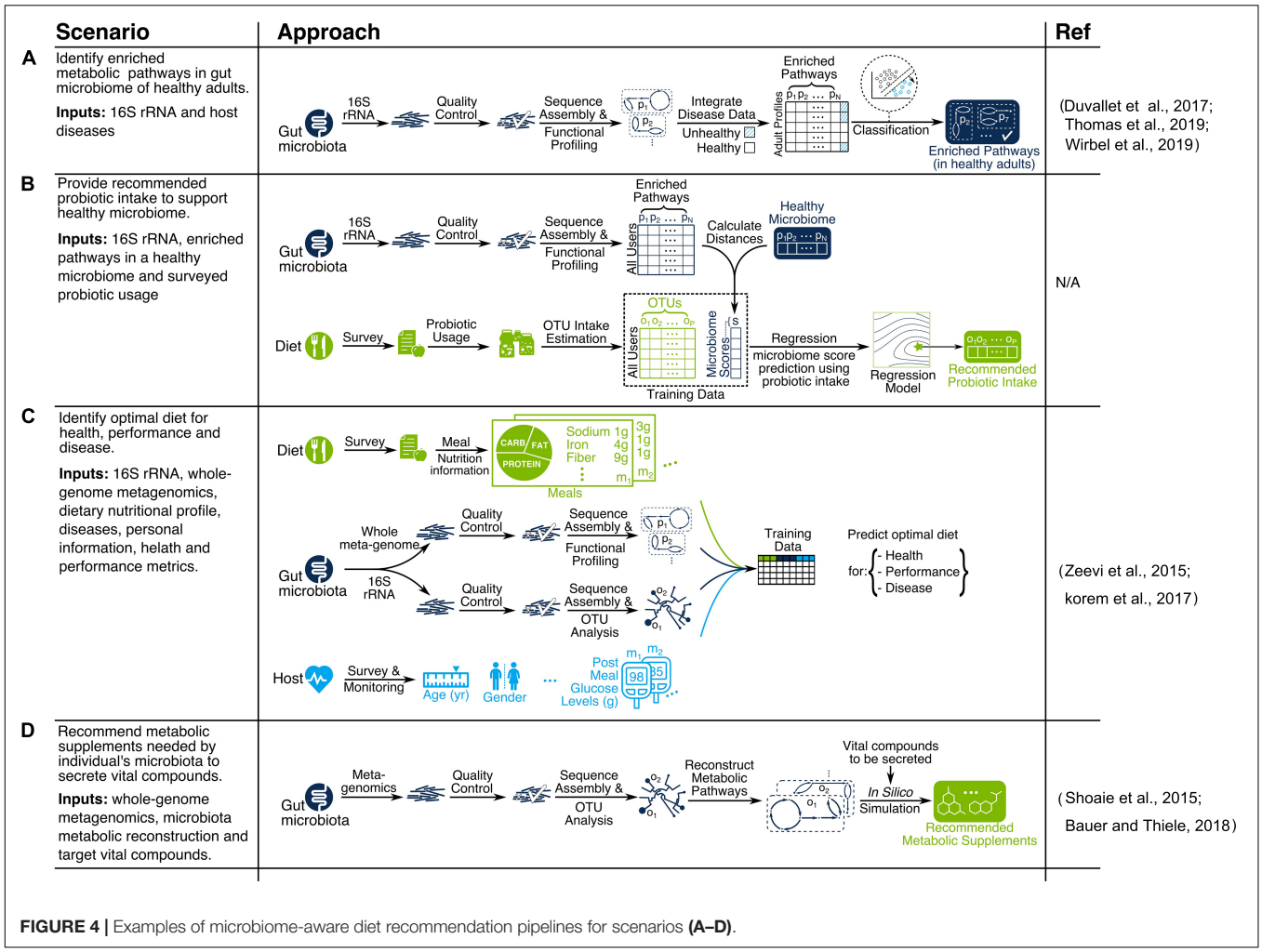


FIGURE 4 | Examples of microbiome-aware diet recommendation pipelines for scenarios **(A–D)**.

Knowledge-Based Recommendation Systems

An ideal knowledge-based RS would be based on *in silico* models that can correctly simulate an individual's gut. It requires proper characterization of the gut microbiome, human intestinal cells, intestinal and dietary metabolite concentrations, their interactions through metabolic pathways, and realistic objective functions for modeling such complex dynamics. Such a knowledge-based RS was devised in a recent study involving 28 patients with Crohn's disease and 26 healthy individuals (Bauer and Thiele, 2018). Researchers integrated genome-scale metabolomic reconstructions (GENREs) of 818 microbes from http://vmh.life (Noronha et al., 2018) with the individual's microbiome abundances after metagenomic data processing in the R package BacArena (Bauer et al., 2017). Their *in silico* simulations provide personalized metabolic supplements for improving patient's SCFA levels. Earlier studies have created a metabolic model of the gut microbiome on a smaller scale (Shoaie et al., 2015). See Magnúsdóttir and Thiele (2018) for a comprehensive review. Despite their promise, there are several challenges for the application of such knowledge-based RSs. The first challenge is the limited availability and accuracy of GENREs for gut microbes. A recent study has identified 1,520 unique microbes in the human gut (Zou et al., 2019), while the number of microbes that have GENREs is only 818 (Noronha et al., 2018). In one study (Tramontano et al., 2018), 75% of the GENREs required updates [from previously constructed GENREs (Magnúsdóttir et al., 2017)] so that *in silico* simulations could recapitulate growth on new media. This suggests that *in silico* GENREs of the gut microbiome are far from complete, however, progress is being made toward closing this gap. The second challenge is the metabolic characterization of the media inside the intestine on which gut microbes grow. This includes identifying the dietary metabolites available to microbes at different sites in the gut, which necessitates meticulous dietary data processing. The third challenge relates to the computational complexity of *in silico* simulations, which increases as host and microbial GENREs become more comprehensive. Although more challenges can be enumerated, their inclusion here would go beyond the scope of this article.

Content-Based Recommendation Systems

In content-based RSs, the recommendations are made based on the item's content (often characterized using item features). This is in contrast to collaborative filtering RSs where

recommendations are based on preferences of other users for each item. In one landmark study (Zeevi et al., 2015), authors use a content-based RS for meal recommendations with the goal of improving post-meal glucose levels. Each meal is first characterized based on its nutritional profile (macronutrients and micronutrients). Then a regression model is trained to predict post-meal glucose level based on the meal's nutritional profile, the individual's microbiome features, and other personal information. For each new user and meal, post-meal glucose levels are predicted by the model, and the meal with the minimum post-meal glucose level is recommended to the user. The same methodology is used in a later study using only microbiome features of individuals to predict post-meal glucose levels in a bread-type recommendation system (Korem et al., 2017). Several challenges arise when building content-based RSs. The first challenge is variable data quality and compatibility. When a group of users (or items) are overrepresented in the data, the predictive model tends to be biased toward their favorite items. As a result, the quality of recommendations will be highly variable. Stratified sampling can be used to alleviate this issue. The second challenge is difficulty in generalizing and personalizing recommendations, particularly when feature vectors are not informative for predictions (also relevant to the "missing quantities" challenge mentioned in **Table 3**). This is in contrast to collaborative filtering RSs, where latent features are learned instead of being defined *a priori*. Hybrid RS methods are designed to take advantage of collaborative filtering RSs to address such inherent challenges in context-based RSs (and vice versa) (Burke, 2002). For an extensive review of context-based RSs, methods see (Lops et al., 2011).

Collaborative Filtering Recommendation Systems

In collaborative filtering RSs, each user is characterized by the items (foods or ingredients here) they have previously rated, bought, or generally acted upon. Recommendations are given based on the idea that users who assign the same rating to existing items are expected to have a similar rating profile for all items. Matrix completion is one of the most popular collaborative filtering methods (Su and Khoshgoftaar, 2009; Ekstrand et al., 2011). User-assigned scores are first organized in a sparse matrix where columns correspond to different items and rows to various users. In cases where most users only have evaluated a few items, most of the matrix remains empty. Matrix completion fills the rest of the matrix through the similarities discovered amongst users and items. See Su and Khoshgoftaar (2009) and Ekstrand et al. (2011) for a comprehensive review. Collaborative filtering RSs have not been used for microbiome-aware food recommendations. We describe an example here to showcase how it can be used. Consider a matrix where each column corresponds to a dietary plan and each row to a person—a specific value can represent gut microbiome alpha diversity during the time which the user followed a particular dietary plan. Assuming that each person has only tried a few dietary plans, most of the matrix will be empty. Here we can use matrix completion to fill the matrix with predicted alpha diversities to create a complete matrix. This can be used to recommend dietary plans for a person with the goal of maximizing gut microbiota diversity. Several challenges arise in collaborative filtering RS. The first challenge is the lack of data for new users ("cold-start"). Note that the recommendations rely on similarities among users, while new users have not tried any of the items available in the database. The second challenge is the curse of dimensionality. As the number of items increases, the chance of having user scores for the same item combinations decreases, hence items and users become equally dissimilar (also relevant to the "high dimensionality" challenges in **Table 3**). In such cases, hybrid RS can be used. Next, we bring up a few example scenarios.

Example Scenarios

We discussed various data analytics and recommendation system methods for microbiome discovery and diet engineering, as illustrated in **Figures 1**, **3**. Applicability of each method depends on research objectives and data availability. Here we explain particular scenarios illustrated in **Figure 4** as blueprints for integrating relevant techniques in a single pipeline. In scenario A, the goal is to identify metabolic pathways that are enriched in the gut microbiome of healthy adults using 16S rRNA data (see Duvallet et al., 2017; Thomas et al., 2019; Wirbel et al., 2019 for similar works). In scenario B, the goal is to provide recommended probiotic intake for supporting a healthy gut

TABLE 4 | Highlighted microbiome-aware diet recommendation studies.

Study description	Dietary variables	Metagenomic technology	References
A personalized meal recommendation system uses personal, microbiome and dietary features to select an optimal meal for lowering post-meal glucose levels in patients with type II diabetes.	Micro and macronutrients	16S rRNA and whole metagenomics	Zeevi et al., 2015
Microbiome features enable accurate prediction of an individual's glycemic response to different bread types.	Bread type	16S rRNA and whole metagenomics	Korem et al., 2017
Accurate prediction of weight regain given normal vs. high-fat diet in mice is enabled using a microbiome-based predictor.	Dietary fat	16S rRNA	Thaiss et al., 2016a
Personalized metabolite supplement recommendations for Crohn's disease are made using *in silico* simulation of reconstructed metabolic pathways from gut microbiome (773 microbes).	Metabolic supplements	Whole metagenomics	Bauer and Thiele, 2018
Fecal amino acid levels are predicted given dietary macronutrients through *in silico* simulation of metabolic pathways from gut microbiome (four microbes) and host cells.	Macronutrients	16S rRNA	Shoaie et al., 2015

microbiome. First, the study participants would be profiled based on the probiotic products they consume (each containing specific OTUs) as well as their gut microbiome. Next, microbiome scores will be calculated for each participant based on the distance between enriched pathways of their microbiome and the target healthy microbiome. Then a regression model is trained to predict microbiome scores based on OTU intakes. Finally, the OTU intake concentration that is predicted to have an optimal microbiome score would be used as the recommended probiotic intake. In scenario C, the goal is to identify optimal diets for health, performance, and disease. A compendium needs to be built following a consistent data collection and processing pipeline for study participants. The compendium serves the training data necessary for building machine learning models to predict health metrics such as post-meal glucose level (Zeevi et al., 2015; Korem et al., 2017) or post-dieting weight regain (Thaiss et al., 2016a). The predictive models can then be used as the key part of a recommendation system by identifying the expected impact of a given diet on health for new individuals. In scenario D, the goal is to recommend metabolic supplements needed by an individual's microbiota to secrete vital compounds. First, OTU abundances of each individual are identified using a metagenomic data processing pipeline. Then, individual gut metabolic pathways are reconstructed using online resources such as the Virtual Metabolic Human database (Noronha et al., 2018). Finally, constraint-based reconstruction and analysis (COBRA) tools (Bauer et al., 2017; Baldini et al., 2018) are used to perform *in silico* simulations of GENREs to identify metabolic intake requirements to secrete vital compounds of interest. This mechanistically sound approach has been used in a few recent studies (Shoaie et al., 2015; Bauer and Thiele, 2018).

INTELLECTUAL PROPERTY DEVELOPMENT

The potential application impact generated by research on the relationship between the gut microbiome and diet can be visualized by the abundant number of patent applications on the topic, as well as more generally in the field of microbiome and health research. A search for "gut microbiome" and "diet" returns over 2,500 patents on Google, deposited by universities, institutes, and companies such as MicroBiome, Microbiome Therapeutics, Gutguide, Whole Biome Inc., UBiome, and others, from as early as 2004. However, it is important to note that most of these hits

TABLE 5 | Highlighted patents relating to diet, gut microbiome, and human health.

Patent number	Name	Owner	Year
US20100172874A1	Gut microbiome as a biomarker and therapeutic target for treating obesity or an obesity-related disorder	Washington University in St. Louis	06
WO2007136553A2	Bacterial strains, compositions including same and probiotic use thereof	Benson et al.	06
US20110123501A1	Gut flora and weight management	Nestec S.A.	07
EP2178543B1	*Lactobacillus rhamnosus* and weight control	Nestec S.A.	07
US9371510B2	Probiotic compositions and methods for inducing and supporting weight loss	Brenda E. Moore	07
US9113641B2	Probiotic bacteria and regulation of fat storage	Arla Foods amba	07
EP2296489A1	*Lactobacillus paracasei* and weight control	Nestec S.A.	08
EP2216036A1	*Lactobacillus rhamnosus* NCC4007, a probiotic mixture and weight control	Nestec S.A.	09
WO2010091991A1	*Lactobacillus helveticus* cncm i-4095 and weight control	Arigoni et al.	09
US20100331641A1	Devices for continual monitoring and introduction of gastrointestinal microbes	Gearbox LLC	09
US20160074505A1	Method and System for Targeting the Microbiome to Promote Health and Treat Allergic and Inflammatory Diseases	Kovarik et al.	09
US20120058094A1	Compositions and methods for treating obesity and related disorders by characterizing and restoring mammalian bacterial microbiota	New York University Dow Global Technologies LLC	10
US9040101B2	Method to treat diabetes utilizing a gastrointestinal microbiome modulating composition	MicroBiome Therapeutics LLC	11
US20170348359A1	Method and System for Treating Cancer and Other Age-Related Diseases by Extending the Health span of a Human	Kovarik et al.	11
US20170281091A1	Capsule device and methodology for discovery of gut microbe roles in diseases with origin in gut	Lowell Zane Shuck	12
US20170372027A1	Method and system for microbiome-derived diagnostics and therapeutics for locomotor system conditions	uBiome Inc.	14
US20170286620A1	Method and system for microbiome-derived diagnostics and therapeutics	uBiome Inc.	14
US20190030095A1	Methods and compositions relating to microbial treatment and diagnosis of disorders	Whole Biome Inc.	14
WO2017216820A1	Metagenomic method *for in vitro* diagnosis of gut dysbiosis	Putignani et al.	16
WO2017171563A1	Beta-caseins and cognitive function	Clarke et al.	16
WO2017160711A1	Modulation of the gut microbiome to treat mental disorders or diseases of the central nervous system	Strandwitz et al.	17
US20180318323A1	Compositions and methods for improving gut health	Plexus Worldwide LLC	17

TABLE 6 | Glossary of terms.

Alpha diversity. A measure that quantifies the species diversity in a given sample. It can be calculated by several methods including richness (i.e. the number of unique species) as well as the Shannon index which relies on the relative abundance of unique species.
Beta diversity. A measure that quantifies the difference between species abundances across samples. It can be calculated by several methods including the Jaccard index (i.e. the ratio of shared to total unique species in a pair of samples) as well as the weighted Jaccard index which also considers the number of times each specie is observed.
Classification. A type of supervised learning problem where the dependent variables are categorical.
Cluster analysis. Unsupervised learning methodology to identify groups of similar datapoints automatically.
Collaborative filtering. Recommendation system methodology which relies on similarities amongst user preferences for new recommendations.
Compositional quantities. Dataset attributes that their absolute quantities are only meaningful relative to each other for each sample, and cannot be compared directly across different samples.
Content-based filtering. Recommendation system methodology in which recommendations are made based on the features for both items and users.
Curse of dimensionality. A set of challenges, such as the need of exponentially more samples to train a model and increased computational complexity, that appear when the dimensionality of the data or model increases.
Data imputation. Substitution of missing values in a given dataset.
Diversity metric. Quantitative measure that represents the number of unique entity types (e.g., species) in a community and evenness in their relative population.
Dimensionality. Number of attributes available for each sample in a given dataset. A dataset with relatively few attributes is considered *low-dimensional* while a dataset with many attributes is referred to as *high-dimensional*.
Labeled/unlabeled samples. Samples that have been tagged using particular labels describing the value of a dependent variable are called *labeled*. This is in contrast to *unlabeled* samples for which such labels are unavailable. Note that labels can be categorical or numerical.
Marker gene sequencing. Primer-based strategy (such as 16S rRNA) that targets a specific region of a gene of interest to characterize microbial phylogenies of a sample.
Multiple-hypothesis testing. A problem that arises in tests of statistical significance when applied multiple times using different hypotheses.
Overfitting. A problem that arises in machine learning where parameter values of a model are too closely fit for training data and therefore not useful in practice.
Rarefaction. A bias correction technique used to enable comparison of diversity measures between communities with unequal sample sizes.
Recommendation system. "Any system that guides a user in a personalized way to interesting or useful objects in a large space of possible options or that produces such objects as output." (Burke, 2002)
Regression. Supervised learning tasks in which the dependent variables are numerical.
Regularization. Machine learning technique that dampens the variability of model parameters leading to a less complex model. It is usually used to mitigate overfitting.
Stability metric. A quantitative measure to assess whether properties of a community (e.g., gut microbes) are preserved over time.
Supervised learning. Learning tasks that require labeled data. They involve learning a function to predict the correct label for a new sample given input attributes.
Unsupervised learning. Learning tasks that do not rely on labeled data. They involve learning hidden structures, features, or patterns within the data.
Variation analysis. Statistical methods, such as analysis of variance (ANOVA), used to identify the amount of variance in a dependent variable that can be explained using independent variables.
Whole metagenomic sequencing. A sequencing strategy that targets the whole genome of all microbial species within a sample. This is also called shotgun metagenomics.

are less than a decade old, demonstrating the relatively early stages in which this area still resides. The exponential growth in patent applications related to the microbiome since 2007 correlates to a similar curve for the academic publications in the same period (Fankhauser et al., 2018).

One of the earliest available patent applications (US20050239706A1) related to the topic of the microbiome and nutrition describes methods to regulate weight by manipulating the gut microbiome. Additional patents also aim to use the gut microbiome as a therapeutic target, monitoring and altering the composition with the goal of manipulating the host phenotype such as weight gain/loss and obesity. In general, weight management with the manipulation of the gut microbiome (US20110123501A1 and US20100172874A1) appears as a favored theme for early patent applications in the area of microbiome and diet. Several patents describe novel probiotics and their uses (WO2007136553A2), often relating them to specific target phenotypes such as weight loss (EP2178543B1, US9371510B2, US9113641B2, EP2216036A1, EP2296489A1, and WO2010091991A1). Multiple applications for probiotics focused on weight loss were deposited by Nestec SA, which offers research and consulting services to the food company Nestlé S.A.

With the development of computational techniques to analyze larger datasets, and more research on the relationship of the microbiome and the host homeostasis and disease, patent applications related to gut microbiome and diet have subsequently extended to other health conditions beyond obesity and weight control. Among the newest patent applications related to the gut microbiome and diet is a patent describing the characterization, diagnostics, and treatment of a locomotor system condition based on microbiome data (US20170372027A1). Other applications include metagenomic methods specific for the comparison of healthy individuals and those with gut dysbiosis (WO2017216820A1), diagnostic tools for Crohn's disease, inflammatory bowel disease, irritable bowel syndrome, ulcerative colitis, and celiac disease using microbiome and other types of data (US20170286620A1), and devices such as capsules to acquire and monitor microbiome and metabolites in the gut (US20170281091A1). Research on the gut–brain axis relationship also resulted in several applications aimed at monitoring and manipulating the gut microbiome to enhance cognition or treat mental-health conditions (WO2017171563A1 and WO2017160711A1). A recent and thorough review of patents related to the microbiome identified cancer diagnosis and treatment and CRISPR technology as recent trends in the field (Fankhauser et al., 2018). **Table 5** shows a summary of highlighted patents relating to DGMH.

Even though there is already a considerable number of patent applications for technologies aiming to manipulate the gut microbiome for multiple health conditions, regulatory legislation has not yet become specific to deal with the new scientific advances in the field. In Europe, the European Food Safety Authority (EFSA) is responsible for regulating and approving food products with health claims, including probiotics, while in the United States, the Food and Drug Administration (FDA) assumes a similar role. Legislation and regulatory aspects are changing in an attempt to keep up with the ever-evolving field. Recently, the FDA has released a statement (Food and Drug Administration, 2018) clarifying existing regulations and announcing the intention to work closely with the United States National Institutes of Health to ensure public safety. Currently, there is no probiotic approved to be marketed in the United States as a live biotherapeutic product, defined by the agency as a "biological product other than a vaccine that contains live organisms used to prevent or treat a disease or condition in humans" (Food and Drug Administration, 2016, 2018). This means that, even though probiotics are legally available as dietary supplements or food ingredients, they cannot yet have claims to cure, treat, or prevent any diseases per current regulation (Food and Drug Administration, 2018), since those claims are reserved for drugs. Classification of food ingredients targeting the microbiome, but not composed of living organisms, microbiota-directed foods or MDFs, prebiotics, and dietary fiber, is also challenging based on the available legislation. Depending on the health claims, such products can fall under the categories of drugs or dietary supplements, which have different requirements for approval (Green et al., 2017).

CONCLUSION

Significant advances in microbiology, genomics, analytical chemistry, computational science, bioinformatics, and other critical disciplines have begun to converge such that it is possible to foresee a new era of health and nutrition research enabling the design of food products capable of optimizing health via predictable interactions with the gut microbiome. Despite the exciting potential in this context demonstrated by pioneering research efforts of many investigators, including those cited in this brief review, the complexity of the microbiome, the chemical composition of food, and their interplay *in situ* remains a daunting challenge in the context of achieving necessary breakthroughs. However, recent advances in high-throughput sequencing and metabolomics profiling, compositional analysis of food, and the emergence of electronic health records as an opportunity to integrate health information provide new sources of data that can contribute to addressing this challenge. Indeed, it is now clear that computational science will play an essential role in this effort as it will provide the foundation to integrate these data layers and derive insights capable of revealing and understanding the complex interactions between diet, microbiome, and health.

The human microbiome is exceptionally plastic, which presents both challenges and opportunities (Gentile and Weir, 2018). Due to its temporal and inter-individual variability, it is difficult to discover statistically significant signatures that unambiguously constitute a healthy versus non-healthy microbiota. At the same time, its potential for adaptation to diet and other environmental factors makes the gut microbiome an excellent target for diet-related interventions to improve health. In this article, we presented a brief overview of the current state of knowledge and potential avenues for research at the interface of diet, gut microbiome, and human health, with particular emphasis on the role that computational science and data analytics can play in accelerating this research. Using these tools, we envision a future in which diets, as well as food and dietary supplement products, can be better designed for specific populations, and, in some cases, for individuals, in order to optimize gut microbiota and health via a platform integrating two distinct systems. The first system will be responsible for identifying the optimal target microbiota (*discovery*) given the desired target, individual, and environment, while the second will provide recommendations for achieving that target microbiota (*engineering*). Recognizing this distinction and the requirement for seamless interaction between the two can reinforce collaborative research in this evolving field where some teams focus on microbiota discovery and others on diet engineering.

Microbiome research has attracted much interest in the past few years and given rise to various software tools and pipelines for metagenomic data processing and analysis. Many of these tools address similar problems and researchers may choose a variety of tools depending on the context. Interestingly, recent research has shown that synthetic datasets can be used to assess the performance of competing tools given a project's assumptions and hence provide useful benchmarks (Ounit and Lonardi, 2016; Hitch and Creevey, 2018). We further believe that progress in simulation-based studies can give rise to new data processing and analytics pipelines customized for each project based on factors such as sequencing technology, data availability, dimensionality, and variability. This can help to build standard protocols for addressing challenges like the ones mentioned in **Tables 3**, **4**.

Our current knowledge about the relationship between diet, gut microbiome, and human health is evolving fast. Many data analysis methods exist for discovering characteristics that can define a healthy microbiota and the factors influencing it. We believe that proper integration of recommendation systems with existing research developments will have an unprecedented impact on our way of life. Given the accelerated pace of advances in sequencing and computational tools, we expect the next decade to be the era of computational nutrition that will revolutionize our relationship with food and diet.

AUTHOR CONTRIBUTIONS

AE, NR, BP, MK, HS, and IT wrote the manuscript. AE and MK created figures with input from all authors. IT supervised

all aspects of the work. All authors reviewed, revised, and approved the manuscript.

FUNDING

The authors acknowledge an unrestricted gift from Mars Inc. to IT, and an NSF SBIR award to PIPA LLC. The funders were not involved in the study design, collection, analysis, and interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

We would like to thank Polina Hadjipanagiotou and the Tagkopoulos lab for their comments on the article.

REFERENCES

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdiscipl. Rev. Comput. Statist.* 5, 149–179. doi: 10.1002/wics.1246

Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L., and Jenmalm, M. C. (2012). Low diversity of the gut microbiota in infants with atopic eczema. *J. Allergy Clin. Immunol.* 129, 434–440.e1-2.

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* 8:e1002358. doi: 10.1371/journal.pcbi.1002358

Agarwala, R., and Morgulis, A. (2011). *BMTagger: Best Match Tagger for Removing Human Reads from Metagenomics Datasets*. Available at: https://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/ (accessed March 14, 2020).

Amann, R. I., Ludwig, W., and Schleifer, K.-H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Mol. Biol. Rev.* 59, 143–169. doi: 10.1128/mmbr.59.1.143-169.1995

American Gut, (2019). *What's in Your Gut?* Available at: http://americangut.org (accessed February 11, 2019).

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e0191-16. doi: 10.1128/mSystems.00191-16

Anderson, J., Edney, R., and Whelan, K. (2012). Systematic review: faecal microbiota transplantation in the management of inflammatory bowel disease. *Aliment. Pharmacol. Therap.* 36, 503–516. doi: 10.1111/j.1365-2036.2012.05220.x

Anderson, M. J. A. (2001). new method for non-parametric multivariate analysis of variance. *Austr. Ecol.* 26, 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Antonazzo, G., Attrill, H., Brown, N., Marygold, S. J., McQuilton, P., Ponting, L., et al. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473:174.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microb.* 17, 690–703. doi: 10.1016/j.chom.2015.04.004

Baldini, F., Heinken, A. K., Heirendt, L., Magnusdottir, S., Fleming, R., and Thiele, I. (2018). The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities. *Bioinformatics* 35, 2332–2334. doi: 10.1093/bioinformatics/bty941

Barabási, A.-L., Menichetti, G., and Loscalzo, J. (2019). The unmapped chemical complexity of our diet. *Nat. Food* 1, 33–37. doi: 10.1038/s43016-019-0005-1

Barratt, M. J., Lebrilla, C., Shapiro, H.-Y., and Gordon, J. I. (2017). The gut microbiota, food science, and human nutrition: a timely marriage. *Cell Host Microb* 22, 134–141. doi: 10.1016/j.chom.2017.07.006

Bartenhagen, C., Klein, H.-U., Ruckert, C., Jiang, X., and Dugas, M. (2010). Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinform.* 11:567. doi: 10.1186/1471-2105-11-567

Bauer, E., and Thiele, I. (2018). From metagenomic data to personalized in silico microbiotas: predicting dietary supplements for Crohn's disease. *NPJ Syst. Biol. Appl.* 4:27.

Bauer, E., Zimmermann, J., Baldini, F., Thiele, I., and Kaleta, C. (2017). BacArena: individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLoS Comput. Biol.* 13:e1005544. doi: 10.1371/journal.pcbi.1005544

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

BioLINCC, (2019). *The Lung HIV Microbiome Project (LHMP)*. Available at: https://biolincc.nhlbi.nih.gov/studies/lhmp/ (accessed February 11, 2019).

Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* 16:191. doi: 10.1186/s13059-015-0759-1

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857.

Bonder, M. J., Kurilshikov, A., Tigchelaar, E. F., Mujagic, Z., Imhann, F., Vila, A. V., et al. (2016). The effect of host genetics on the gut microbiome. *Nat. Genet.* 48:1407. doi: 10.1038/ng.3663

Bradley, P. H., Nayfach, S., and Pollard, K. S. (2018). Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLoS Computat. Biol.* 14:e1006242. doi: 10.1371/journal.pcbi.1006242

Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P., and Tyson, G. W. (2012). Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat. Methods* 9:425. doi: 10.1038/nmeth.1990

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325–349. doi: 10.2307/1942268

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Breitwieser, F. P., Lu, J., and Salzberg, S. L. A. (2017). review of methods and databases for metagenomic classification and assembly. *Briefings Bioinform.* 20, 1125–1136. doi: 10.1093/bib/bbx120

Brown, C. T., Davis-Richardson, A. G., Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., et al. (2011). Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS One* 6:e25792. doi: 10.1371/journal.pone.0025792

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12:59. doi: 10.1038/nmeth.3176

Burke, R. (2002). Hybrid recommender systems: survey and experiments. *User Model. User Adapt. Interact.* 12, 331–370.

Caliński, T., and Harabasz, J. A. (1974). dendrite method for cluster analysis. *Commun. Statist. Theor. Methods* 3, 1–27. doi: 10.1080/03610927408827101

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13:581. doi: 10.1038/nmeth.3869

Canada, H. (2010). *Canadian Nutrient File*. Ottawa: Government of Canada Ottawa.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods.* 7:335.

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., et al. (2017). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* 46, D633–D639.

Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., et al. (2018). The gut microbiome profile in obesity: a systematic review. *Int. J. Endocrinol.* 2018:9.

Chen, E. Z., and Li, H. A. (2016). two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32, 2611–2617. doi: 10.1093/bioinformatics/btw308

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Austr. J. Ecol.* 18, 117–143. doi: 10.1111/j.1442-9993.1993.tb00438.x

ClinicalTrials.gov, (2019). *Pediatric Metabolism and Microbiome Repository - Full Text View*. Available at: https://clinicaltrials.gov/ct2/show/NCT02959034 (accessed February 11, 2019).

Conlon, M., and Bird, A. (2015). The impact of diet and lifestyle on gut microbiota and human health. *Nutrients* 7, 17–44. doi: 10.3390/nu7010017

Costea, P. I., Hildebrand, F., Manimozhiyan, A., Bäckhed, F., Blaser, M. J., Bushman, F. D., et al. (2018). Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* 3:8. doi: 10.1038/s41564-017-0072-8

Cox, L. M., and Blaser, M. J. (2015). Antibiotics in early life and obesity. *Nat. Rev. Endocrinol.* 11:182. doi: 10.1038/nrendo.2014.210

Coyte, K. Z., Schluter, J., and Foster, K. R. (2015). The ecology of the microbiome: networks, competition, and stability. *Science* 350, 663–666. doi: 10.1126/science.aad2602

curatedMetagenomicData, (2019). *curatedMetagenomicData*. Available at: http://waldronlab.io/curatedMetagenomicData (accessed February 11, 2019).

Data – The Harvard Personal Genome Project (PGP), (2019). Available at: https://pgp.med.harvard.edu/data (accessed February 11, 2019).

Davenport, E. R., Cusanovich, D. A., Michelini, K., Barreiro, L. B., Ober, C., and Gilad, Y. (2015). Genome-wide association studies of the human gut microbiota. *PLoS One* 10:e0140301. doi: 10.1371/journal.pone.0140301

Davenport, E. R., Mizrahi-Man, O., Michelini, K., Barreiro, L. B., Ober, C., and Gilad, Y. (2014). Seasonal variation in human gut microbiome composition. *PLoS One* 9:e90731. doi: 10.1371/journal.pone.0090731

De Filippis, F., Pellegrini, N., Vannini, L., Jeffery, I. B., La Storia, A., Laghi, L., et al. (2016). High-level adherence to a mediterranean diet beneficially impacts the gut microbiota and associated metabolome. *Gut* 65, 1812–1821. doi: 10.1136/gutjnl-2015-309957

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14691–14696. doi: 10.1073/pnas.1005963107

de Goffau, M. C., Luopajärvi, K., Knip, M., Ilonen, J., Ruohtula, T., Härkönen, T., et al. (2013). Fecal microbiota composition differs between children With β-cell autoimmunity and those without. *Diabetes* 62, 1238–1244. doi: 10.2337/db12-0526

de Steenhuijsen Piters, W. A., Heinonen, S., Hasrat, R., Bunsow, E., Smith, B., Suarez-Arrabal, M.-C., et al. (2016). Nasopharyngeal microbiota, host transcriptome, and disease severity in children with respiratory syncytial virus infection. *Am. J. Respir. Crit. Care Med.* 194, 1104–1115. doi: 10.1164/rccm.201602-0220oc

Delzenne, N. M., Olivares, M., Neyrinck, A. M., Beaumont, M., Kjølbæk, L., Larsen, T. M., et al. (2019). Nutritional interest of dietary fiber and prebiotics in obesity: Lessons from the MyNewGut consortium. *Clin. Nutr.* 39, 414–424. doi: 10.1016/j.clnu.2019.03.002

DeSantis, T., Hugenholtz, P., Keller, K., Brodie, E., Larsen, N., Piceno, Y., et al. (2006). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* 2006, W394–W399.

Dobson, A. J., Chaston, J. M., Newell, P. D., Donahue, L., Hermann, S. L., Sannino, D. R., et al. (2015). Host genetic determinants of microbiota-dependent nutrition revealed by genome-wide analysis of Drosophila melanogaster. *Nat. Commun.* 6:6312. doi: 10.1038/ncomms7312

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., et al. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U.S.A.* 107:11971. doi: 10.1073/pnas.1002601107

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* 8:1784. doi: 10.1038/s41467-017-01973-8

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C. (2013). highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10:996. doi: 10.1038/nmeth.2604

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*. 2016:081257.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Eetemadi, A., and Tagkopoulos, I. (2018). Genetic Neural Networks: An artificial neural network architecture for capturing gene expression relationships. *Bioinformatics* 19:bty945. doi: 10.1093/bioinformatics/bty945

Ekstrand, M. D., Riedl, J. T., and Konstan, J. A. (2011). Collaborative filtering recommender systems. *Found. Trends Hum. Comput. Interact.* 4, 81–173.

EMBL-EBI Mg, (2019). *MGnify home page > EMBL-EBI [Internet]. MGnify*. Available at: https://www.ebi.ac.uk/metagenomics/ (accessed February 11, 2019)

European Nucleotide Archive EMBL-EBI, (2019). Available from: https://www.ebi.ac.uk/ena (accessed February 11, 2019).

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2017). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46, D649–D655.

Fankhauser, M., Moser, C., and Nyfeler, T. (2018). Patents as early indicators of technology and investment trends: analyzing the microbiome space as a case study. *Front. Bioeng. Biotechnol.* 6:84. doi: 10.3389/fbioe.2018.00084

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am. Naturalist.* 125, 1–15. doi: 10.1086/284325

Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., et al. (2018). Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microb.* 24, 133–145. doi: 10.1016/j.chom.2018.06.005

Food and Drug Administration (2016). *Early Clinical Trials with Live Biotherapeutic Products: Chemistry, Manufacturing, and Control Information*. Available at: https://www.fda.gov/downloads/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/General/UCM292704.pdf

Food and Drug Administration (2018). *Statement from FDA Commissioner Scott Gottlieb, M.D., on Advancing The Science and Regulation of Live Microbiome-Based Products Used to Prevent, Treat, or Cure Diseases in Humans*. Available at: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm617168.htm

Foster, K. R., Schluter, J., Coyte, K. Z., and Rakoff-Nahoum, S. (2017). The evolution of the host microbiome as an ecosystem on a leash. *Nature* 548:43. doi: 10.1038/nature23292

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47:1091. doi: 10.1038/ng.3367

Gentile, C. L., and Weir, T. L. (2018). The gut microbiota at the intersection of diet and human health. *Science* 362, 776–780. doi: 10.1126/science.aau5812

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.* 24:392. doi: 10.1038/nm.4517

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Goldenberg, J. Z., Yap, C., Lytvyn, L., Lo, C. K.-F., Beardsley, J., Mertz, D., et al. (2017). Probiotics for the prevention of clostridium difficile-associated diarrhea in adults and children. *Cochrane Database Syst. Rev.* 112:CD006095.

Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., et al. (2016). Genetic determinants of the gut microbiome in UK twins. *Cell Host Microb.* 19, 731–743. doi: 10.1016/j.chom.2016.04.017

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., et al. (2014). Human genetics shape the gut microbiome. *Cell* 159, 789–799. doi: 10.1016/j.cell.2014.09.053

Gopalakrishnan, V., Spencer, C., Nezi, L., Reuben, A., Andrews, M., Karpinets, T., et al. (2018). Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science* 359, 97–103.

Gould, A. L., Zhang, V., Lamberti, L., Jones, E. W., Obadia, B., Korasidis, N., et al. (2018). Microbiome interactions shape host fitness. *Proc. Natl. Acad. Sci. U.S.A.* 115, E11951–E11960.

Green, J. M., Barratt, M. J., Kinch, M., and Gordon, J. I. (2017). Food and microbiota in the FDA regulatory framework. *Science* 357, 39–40. doi: 10.1126/science.aan0836

Greenhalgh, K., Ramiro-Garcia, J., Heinken, A., Ullmann, P., Bintener, T., Pacheco, M. P., et al. (2018). Integrated in vitro and in silico modelling delineates the molecular effects of a symbiotic regimen on colorectal cancer-derived cells. *Cell Rep.* 27, 1621–1632.e9. doi: 10.1016/j.celrep.2019.04.001

Grice, E. A., and Segre, J. A. (2012). The human microbiome: our second genome. *Annu. Rev. Genom. Hum. Genet.* 13, 151–170. doi: 10.1146/annurev-genom-090711-163814

Guo, Y., Hastie, T., and Tibshirani, R. (2006). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 8, 86–100. doi: 10.1093/biostatistics/kxj035

Hall, A. B., Tolonen, A. C., and Xavier, R. J. (2017). Human genetic variation and the gut microbiome in disease. *Nat. Rev. Genet.* 18:690. doi: 10.1038/nrg.2017.63

Hao, Q., Dong, B. R., and Wu, T. (2015). Probiotics for preventing acute upper respiratory tract infections. *Cochrane Database Syst. Rev.* 2:CD006895.

Heintz-Buschart, A., and Wilmes, P. (2017). Human gut microbiome: function matters. *Trends Microbiol.* 26, 563–574. doi: 10.1016/j.tim.2017.11.002

Hildebrand, F., Nguyen, T. L. A., Brinkman, B., Yunta, R. G., Cauwe, B., Vandenabeele, P., et al. (2013). Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol.* 14:R4. doi: 10.1186/gb-2013-14-1-r4

Hitch, T. C., and Creevey, C. J. (2018). Spherical: an iterative workflow for assembling metagenomic datasets. *BMC Bioinformatics* 19:20. doi: 10.1186/s12859-018-2028-2

HOMD : Human Oral Microbiome Database, (2019). Available at: http://www.homd.org/index.php (accessed February 11, 2019).

Hotelling, H. (1992). "Relations between two sets of variates," in *Breakthroughs in statistics*, eds S. Kotz, and N. L. Johnson, (New York, NY: Springer), 162–190. doi: 10.1007/978-1-4612-4380-9_14

Howie, B. N., Donnelly, P., and Marchini, J. A. (2009). flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

HPMCD: Human Pan Microbial Communities Database, (2019). Available at: http://www.hpmcd.org (accessed February 11, 2019).

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8:84. doi: 10.3389/fgene.2017.00084

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* 12:e1004957. doi: 10.1371/journal.pcbi.1004957

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486:207. doi: 10.1038/nature11234

IBM Corp, (2013). *IBM SPSS Statistics For Windows. Version 220*. Armonk, NY: IBM Corp.

Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J., et al. (2019). Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. *Front. Genet.* 10:995.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114.

Karczewski, K. J., and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19:299. doi: 10.1038/nrg.2018.4

Katoh, K., and Standley, D. M. M. A. F. F. T. (2013). multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evolut.* 30, 772–780. doi: 10.1093/molbev/mst010

Kaufman, L., and Rousseeuw, P. (1987). *Clustering by Means Of Medoids*. North-Holland: Delft university.

Kim, M., and Tagkopoulos, I. (2018). Data integration and predictive modeling methods for multi-omics datasets. *Mol. Omics.* 14, 8–25. doi: 10.1039/c7mo00051k

Kim, M., Eetemadi, A., and Tagkopoulos, I. (2017). DeepPep: deep proteome inference from peptide profiles. *PLoS Comput. Biol.* 13:e1005661. doi: 10.1371/journal.pcbi.1005661

Kim, M., Rai, N., Zorraquino, V., and Tagkopoulos, I. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nat. Commun.* 7:13090. doi: 10.1038/ncomms13090

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16, 410–422.

Korem, T., Zeevi, D., Zmora, N., Weissbrod, O., Bar, N., Lotan-Pompan, M., et al. (2017). Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. *Cell Metab.* 25, 1243–1253. doi: 10.1016/j.cmet.2017.05.002

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., et al. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* 9:e1002863. doi: 10.1371/journal.pcbi.1002863

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microb.* 17, 260–273. doi: 10.1016/j.chom.2015.01.001

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1–27. doi: 10.1007/bf02289565

Kultima, J. R., Coelho, L. P., Forslund, K., Huerta-Cepas, J., Li, S. S., Driessen, M., et al. (2016). MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* 32, 2520–2523. doi: 10.1093/bioinformatics/btw183

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31:814. doi: 10.1038/nbt.2676

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357. doi: 10.1038/nmeth.1923

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Statist. Softw.* 25, 1–18.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. M. E. G. A. H. I. T. (2015). an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Ann. Rev. Statist. Appl.* 2, 73–94. doi: 10.1146/annurev-statistics-010814-020351

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32:834. doi: 10.1038/nbt.2942

Lops, P., De Gemmis, M., and Semeraro, G. (2011). *Content-Based Recommender Systems: State Of The Art And Trends. In: Recommender Systems Handbook*. Berlin: Springer, 73–105.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 489:220. doi: 10.1038/nature11550

Magnúsdóttir, S., and Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Curr. Biotechnol.* 51, 90–96. doi: 10.1016/j.copbio.2017.12.005

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., et al. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* 35:81. doi: 10.1038/nbt.3703

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., et al. (2006). Reduced diversity of faecal microbiota in Crohn's disease

revealed by a metagenomic approach. *Gut* 55, 205–211. doi: 10.1136/gut.2005.073817

Mansfield, J. A., Bergin, S. W., Cooper, J. R., and Olsen, C. H. (2014). Comparative probiotic strain efficacy in the prevention of eczema in infants and children: a systematic review and meta-analysis. *Mil. Med.* 179, 580–592. doi: 10.7205/MILMED-D-13-00546

Martiny, A. C., Martiny, J. B. H., Weihe, C., Field, A., and Ellis, J. (2011). Functional metagenomics reveals previously unrecognized diversity of antibiotic resistance genes in gulls. *Front. Microbiol.* 2:238. doi: 10.3389/fmicb.2011.00238

Mathew, S., Smatti, M. K., Al Ansari, K., Nasrallah, G. K., Al Thani, A. A., and Yassine, H. M. (2019). Mixed viral-bacterial infections and their effects on gut microbiota and clinical illnesses in children. *Sci. Rep.* 9:865. doi: 10.1038/s41598-018-37162-w

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6:610. doi: 10.1038/ismej.2011.139

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Mehta, R. S., Abu-Ali, G. S., Drew, D. A., Lloyd-Price, J., Subramanian, A., Lochhead, P., et al. (2018). Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* 3:347. doi: 10.1038/s41564-017-0096-0

Menni, C., Lin, C., Cecelja, M., Mangino, M., Matey-Hernandez, M. L., Keehn, L., et al. (2018). Gut microbial diversity is associated with lower arterial stiffness in women. *Eur. Heart J.* 39, 2390–2397. doi: 10.1093/eurheartj/ehy226

MG-RAST, (2019). Available at: http://www.mg-rast.org (accessed February 11, 2019).

Moossavi, S., Sepehri, S., Robertson, B., Bode, L., Goruk, S., Field, C. J., et al. (2019). Composition and variation of the human milk microbiota are influenced by maternal and early-life factors. *Cell Host Microb.* 25, 324–335. doi: 10.1016/j.chom.2019.01.011

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 13:R79.

Murtagh, F., and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdiscipl. Rev. Data Min. Knowl. Discov.* 2, 86–97.

NASA, (2019). *Study of the Impact of Long-Term Space Travel on the Astronauts' Microbiome*. Available at: https://www.nasa.gov/mission_pages/station/research/experiments/explorer/Investigation.html?#id=982 (accessed February 11, 2019).

NIH Human Microbiome Project - Home, (2019). Available at: https://hmpdacc.org/hmp (accessed February 11, 2019).

Noronha, A., Modamio, J., Jarosz, Y., Guerard, E., Sompairac, N., Preciat, G., et al. (2018). The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Res.* 47, D614–D624.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi: 10.1101/gr.213959.116

Ospina, R., and Ferrari, S. L. A. (2012). general class of zero-or-one inflated beta regression models. *Comput. Statist. Data Analy.* 56, 1609–1623. doi: 10.1016/j.csda.2011.10.005

Ounit, R., and Lonardi, S. (2016). Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics* 32, 3823–3825. doi: 10.1093/bioinformatics/btw542

Pan, S. J., and Yang, Q. A. (2010). Survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 10, 1345–1359.

Pan, W.-H., Sommer, F., Falk-Paulsen, M., Ulas, T., Best, P., Fazio, A., et al. (2018). Exposure to the gut microbiota drives distinct methylome and transcriptome changes in intestinal epithelial cells during postnatal development. *Genome Med.* 10:27. doi: 10.1186/s13073-018-0534-5

Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A., et al. (2017). A microbial signature for Crohn's disease. *Gut* 66, 813–822.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* 14:1023. doi: 10.1038/nmeth.4468

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10:1200. doi: 10.1038/nmeth.2658

Peng, X., Li, G., and Liu, Z. (2016). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J. Comput. Biol.* 23, 102–110. doi: 10.1089/cmb.2015.0157

Pereira, J., Rea, K., Nolan, Y., O'Leary, O., Dinan, T., and Cryan, J. (2019). Depression's unholy trinity: dysregulated stress, immunity, and the microbiome. *Ann. Rev. Psychol.* 71, 49–78. doi: 10.1146/annurev-psych-122216-011613

Prehn-Kristensen, A., Zimmermann, A., Tittmann, L., Lieb, W., Schreiber, S., Baving, L., et al. (2018). Reduced microbiome alpha diversity in young patients with ADHD. *PLoS One* 13:e0200728. doi: 10.1371/journal.pone.0200728

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490:55. doi: 10.1038/nature11450

Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of machine learning in Microbiology. *Front. Microbiol.* 10:827. doi: 10.3389/fmicb.2019.00827

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35:833. doi: 10.1038/nbt.3935

Quinn, R. A., Navas-Molina, J. A., Hyde, E. R., Song, S. J., Vázquez-Baeza, Y., Humphrey, G., et al. (2016). From sample to multi-omics conclusions in under 48 hours. *mSystems* 1:e0038-6.

Rahman, S. F., Olm, M. R., Morowitz, M. J., and Banfield, J. F. (2018). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *mSystems* 3:e00123-17. doi: 10.1128/mSystems.00123-17

Ranjan, R., Rani, A., Finn, P. W., and Perkins, D. L. (2018). Evaluating bacterial and functional diversity of human gut microbiota by complementary metagenomics and metatranscriptomics. *bioRxiv* [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/363200v1

Raymond, F., Ouameur, A. A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., et al. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.* 10:707. doi: 10.1038/ismej.2015.148

Relman, D. A. (2012). The human microbiome: ecosystem resilience and health. *Nutr. Rev.* 70(Suppl._1), S2–S9.

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution* 43, 223–225. doi: 10.1111/j.1558-5646.1989.tb04220.x

Robertson, R. C., Kaliannan, K., Strain, C. R., Ross, R. P., Stanton, C., and Kang, J. X. (2018). Maternal omega-3 fatty acids regulate offspring obesity through persistent modulation of gut microbiota. *Microbiome* 6:95. doi: 10.1186/s40168-018-0476-6

Rodríguez, J. M., Murphy, K., Stanton, C., Ross, R. P., Kober, O. I., Juge, N., et al. (2015). The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb. Ecol. Health Dis.* 26:10.3402/mehd.v26.26050. doi: 10.3402/mehd.v26.26050

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. V. (2016). a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling and more. Version 0.5–12 (BETA). *J. Statist. Softw.* 48, 1–36.

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature* 555:210. doi: 10.1038/nature25973

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Ruan, Q., Dutta, D., Schwalbach, M. S., Steele, J. A., Fuhrman, J. A., and Sun, F. (2006). Local similarity analysis reveals unique associations among marine

bacterioplankton species and environmental factors. *Bioinformatics* 22, 2532–2538. doi: 10.1093/bioinformatics/btl417

Saez-Lara, M. J., Gomez-Llorente, C., Plaza-Diaz, J., and Gil, A. (2015). The role of probiotic lactic acid bacteria and bifidobacteria in the prevention and treatment of inflammatory bowel disease and other related diseases: a systematic review of randomized human clinical trials. *Biomed. Res. Int.* 2015:505878. doi: 10.1155/2015/505878

Sánchez, B., Delgado, S., Blanco-Míguez, A., Lourenço, A., Gueimonde, M., and Margolles, A. (2017). Probiotics, gut microbiota, and their influence on host health and disease. *Mol. Nutr. Food Res.* 61:1600240. doi: 10.1002/mnfr.201600240

Scher, J. U., Ubeda, C., Artacho, A., Attur, M., Isaac, S., Reddy, S. M., et al. (2015). Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease. *Arthrit. Rheumatol.* 67, 128–139. doi: 10.1002/art.38892

Schippa, S., Iebba, V., Barbato, M., Di Nardo, G., Totino, V., Checchi, M. P., et al. (2010). A distinctive "microbial signature" in celiac pediatric patients. *BMC Microbiol.* 10, 1471–2180.

Schloss, P. D., and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* 71, 1501–1506. doi: 10.1128/aem.71.3.1501-1506.2005

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Schmieder, R., and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* 6:e17288. doi: 10.1371/journal.pone.0017288

Scholkopf, B., and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, And Beyond*. Cambridge, MA: MIT press.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res.* 99, 323–338.

Schwartz, S., Friedberg, I., Ivanov, I. V., Davidson, L. A., Goldsby, J. S., Dahl, D. B., et al. (2012). A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genom. Biol.* 13:r32. doi: 10.1186/gb-2012-13-4-r32

Shao, Y., Forster, S. C., Tsaliki, E., Vervier, K., Strang, A., Simpson, N., et al. (2019). Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* 574, 117–121. doi: 10.1038/s41586-019-1560-1

Shavitt, I., and Segal, E. (2018). Regularization learning networks: deep learning for tabular datasets. *Adv. Neural Inform. Process. Syst.* 1384–1394. Available at : http://papers.nips.cc/paper/7412-regularization-learning-networks-deep-learning-for-tabular-datasets

Sherwin, E., Dinan, T. G., and Cryan, J. F. (2018). Recent developments in understanding the role of the gut microbiota in brain health and disease. *Ann. N. Y. Acad. Sci.* 1420, 5–25. doi: 10.1111/nyas.13416

Shim, J.-S., Oh, K., and Kim, H. C. (2014). Dietary assessment methods in epidemiologic studies. *Epidemiol. Health* 36:e2014009. doi: 10.4178/epih/e2014009

Shoaie, S., Ghaffari, P., Kovatcheva-Datchary, P., Mardinoglu, A., Sen, P., Pujos-Guillot, E., et al. (2015). Quantifying diet-induced metabolic changes of the human gut microbiome. *Cell Metab.* 22, 320–331. doi: 10.1016/j.cmet.2015.07.001

Shreiner, A. B., Kao, J. Y., and Young, V. B. (2015). The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* 31, 69–75. doi: 10.1097/MOG.0000000000000139

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. A. (2017). phylogenetic transform enhances analysis of compositional microbiota data. *eLife* 6:e21887. doi: 10.7554/eLife.21887

Singh, R., Lanchantin, J., Robins, G., and Qi, Y. (2016). Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 32, i639–i648. doi: 10.1093/bioinformatics/btw427

Smith, E. P., and van Belle, G. (1984). Nonparametric estimation of species richness. *Biometrics* 40, 119–129.

Smith, H., Gnanadesikan, R., and Hughes, J. B. (1962). Multivariate analysis of variance (MANOVA). *Biometrics* 18, 22–41.

Sordillo, J. E., Korrick, S., Laranjo, N., Carey, V., Weinstock, G. M., Gold, D. R., et al. (2019). Association of the infant gut microbiome with early childhood neurodevelopmental outcomes: an ancillary study to the VDAART randomized clinical trial. *JAMA Netw. Open* 2:e190905. doi: 10.1001/jamanetworkopen.2019.0905

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101.

Stinson, L. F., Boyce, M. C., Payne, M. S., and Keelan, J. A. (2019). The not-so-sterile womb: Evidence that the human fetus is exposed to bacteria prior to birth. *Front. Microbiol.* 10:1124. doi: 10.3389/fmicb.2019.01124

Su, X., and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. Artif. Intellig.* 2009:421425.

Suphavilai, C., Bertrand, D., and Nagarajan, N. (2018). Predicting cancer drug response using a recommender system. *Bioinformatics* 34, 3907–3914. doi: 10.1093/bioinformatics/bty452

Suykens, J. A., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9, 293–300.

Taroncher-Oldenburg, G., Jones, S., Blaser, M., Bonneau, R., Christey, P., Clemente, J. C., et al. (2018). *Translating Microbiome Futures*. London: Nxature Publishing Group.

Thaiss, C. A., Itav, S., Rothschild, D., Meijer, M. T., Levy, M., Moresi, C., et al. (2016a). Persistent microbiome alterations modulate the rate of post-dieting weight regain. *Nature* 540:544. doi: 10.1038/nature20796

Thaiss, C. A., Levy, M., Korem, T., Dohnalová, L., Shapiro, H., Jaitin, D. A., et al. (2016b). Microbiota diurnal rhythmicity programs host transcriptome oscillations. *Cell* 167, 1495–1510. doi: 10.1016/j.cell.2016.11.003

Thaiss, C. A., Zeevi, D., Levy, M., Zilberman-Schapira, G., Suez, J., Tengeler, A. C., et al. (2014). Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell* 159, 514–529. doi: 10.1016/j.cell.2014.09.048

Thavagnanam, S., Fleming, J., Bromley, A., Shields, M. D., and Cardwell, C. R. A. (2008). meta-analysis of the association between Caesarean section and childhood asthma. *Clin. Exp. Allergy* 38, 629–633. doi: 10.1111/j.1365-2222.2007.02780.x

The Michigan Microbiome Project, (2019). *University of Michigan | Center for Microbial Systems*. Available at: https://microbe.med.umich.edu/research/michigan-microbiome-project (accessed February 11, 2019).

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7

Tibshirani, R., and Walther, G. (2005). Cluster validation by prediction strength. *J. Comput. Graph. Statist.* 14, 511–528. doi: 10.1198/106186005x59243

Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M., Wiens, J., and Schloss, P. D. (2019). Effective application of machine learning to microbiome-based classification problems. *BioRxiv* [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/816090v1 (accessed March 14, 2020).

Tramontano, M., Andrejev, S., Pruteanu, M., Klünemann, M., Kuhn, M., Galardini, M., et al. (2018). Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. *Nat. Microbiol.* 3:514. doi: 10.1038/s41564-018-0123-9

Tran, T. N. T., Atas, M., Felfernig, A., and Stettinger, M. (2017). An overview of recommender systems in the healthy food domain. *J. Intellig. Inform. Syst.* 50, 501–526. doi: 10.1007/s10844-017-0469-0

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12:902. doi: 10.1038/nmeth.3589

Tsilimigras, M. C., and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* 26, 330–335. doi: 10.1016/j.annepidem.2016.03.002

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457:480. doi: 10.1038/nature07540

Turpin, W., Espin-Garcia, O., Xu, W., Silverberg, M. S., Kevans, D., Smith, M. I., et al. (2016). Association of host genome with intestinal microbial

composition in a large healthy cohort. *Nat. Genet.* 48:1413. doi: 10.1038/ng.3693

TwinsUK, (2019). *The Biggest twin Registry In The Uk for The Study Of Ageing Related Diseases*. Available at: http://twinsuk.ac.uk (accessed February 11, 2019)

Tyler, A. D., Smith, M. I., and Silverberg, M. S. (2014). Analyzing the human microbiome: a "how to" guide for physicians. *Am. J. Gastroenterol.* 109:983. doi: 10.1038/ajg.2014.73

US Department of Agriculture and Agricultural Research Service, (2010). *USDA National Nutrient Database for Standard Reference, Release 28. Agricultural Research Service*. Washington, D.C: USDA.

Vatanen, T., Franzosa, E. A., Schwager, R., Tripathi, S., Arthur, T. D., Vehik, K., et al. (2018). The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* 562:589. doi: 10.1038/s41586-018-0620-2

Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A., Rosenquist, M., Halfvarson, J., et al. (2009). Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* 3:179. doi: 10.1038/ismej.2008.108

Walker, A., Pfitzner, B., Neschen, S., Kahle, M., Harir, M., Lucio, M., et al. (2014). Distinct signatures of host–microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. *ISME J.* 8:2380. doi: 10.1038/ismej.2014.79

Wang, J., Kurilshikov, A., Radjabzadeh, D., Turpin, W., Croitoru, K., Bonder, M. J., et al. (2018). Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. *Biomed. Central* 6:101.

Wang, M., Karlsson, C., Olsson, C., Adlerberth, I., Wold, A. E., Strachan, D. P., et al. (2008). Reduced diversity in the early fecal microbiota of infants with atopic eczema. *J. Allergy Clin. Immunol.* 121, 129–134. doi: 10.1016/j.jaci.2007.09.011

Washburne, A. D., Morton, J. T., Sanders, J., McDonald, D., Zhu, Q., Oliverio, A. M., et al. (2018). Methods for phylogenetic analysis of microbiome data. *Nat. Microbiol.* 3:652. doi: 10.1038/s41564-018-0156-0

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10:1669. doi: 10.1038/ismej.2015.235

Willing, B. P., Russell, S. L., and Finlay, B. B. (2011). Shifting the balance: antibiotic effects on host–microbiota mutualism. *Nat. Rev. Microbiol.* 9:233. doi: 10.1038/nrmicro2536

Willis, A. D. (2019). Rarefaction, alpha diversity, and statistics. *Front. Microbiol.* 10:2407. doi: 10.3389/fmicb.2019.02407

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25, 679–689. doi: 10.1038/s41591-019-0406-6

Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* 15:473. doi: 10.1038/nrd.2016.32

Witten, D. M., Tibshirani, R., and Hastie, T. A. (2009). penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 515–534. doi: 10.1093/biostatistics/kxp008

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genom. Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Wooley, J. C., Godzik, A., and Friedberg, I. A. (2010). primer on metagenomics. *PLoS Comput. Biol.* 6:e1000667. doi: 10.1371/journal.pcbi.1000667

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486:222. doi: 10.1038/nature11053

Yong, A. G., and Pearce, S. A. (2013). beginner's guide to factor analysis: focusing on exploratory factor analysis. *Tutor. Quant. Methods Psychol.* 9, 79–94. doi: 10.20982/tqmp.09.2.p079

Yuan, C., Gaskins, A. J., Blaine, A. I., Zhang, C., Gillman, M. W., Missmer, S. A., et al. (2016). Cesarean birth and risk of offspring obesity in childhood, adolescence and early adulthood. *JAMA Pediatr.* 170:e162385. doi: 10.1001/jamapediatrics.2016.2385

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163, 1079–1094. doi: 10.1016/j.cell.2015.11.001

Zhang, C., Yin, A., Li, H., Wang, R., Wu, G., Shen, J., et al. (2015). Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *eBio Med.* 2, 968–984. doi: 10.1016/j.ebiom.2015.07.007

Zhang, X., Deeke, S. A., Ning, Z., Starr, A. E., Butcher, J., Li, J., et al. (2018). Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.* 9:2873. doi: 10.1038/s41467-018-05357-4

Zhao, L., Zhang, F., Ding, X., Wu, G., Lam, Y. Y., Wang, X., et al. (2018). Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* 359, 1151–1156. doi: 10.1126/science.aao5774

Zheng, H., Liang, H., Wang, Y., Miao, M., Shi, T., Yang, F., et al. (2016). Altered gut microbiota composition associated with eczema in infants. *PLoS One* 11:e0166026. doi: 10.1371/journal.pone.0166026

Zhou, Y.-H., and Gallins, P. A. (2019). review and tutorial of machine learning methods for microbiome host trait prediction. *Front. Genet.* 10:579. doi: 10.3389/fgene.2019.00579

Zhu, X. (2005). *Semi-Supervised Learning Literature Survey*. Wisconsin: University of Wisconsin-Madison.

Zmora, N., Suez, J., and Elinav, E. (2019). You are what you eat: diet, health and the gut microbiota. *Nat. Rev. Gastroenterol. Hepatol.* 16, 35–56. doi: 10.1038/s41575-018-0061-2

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* 37:179. doi: 10.1038/s41587-018-0008-8

Conflict of Interest: MK and IT are employed or have a financial interest in PIPA LLC. HS has a financial interest in T.O.P. LLC and March Capital US LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 15 April 2020
doi: 10.3389/fmicb.2020.00579

Predicting Microbe-Disease Association by Learning Graph Representations and Rule-Based Inference on the Heterogeneous Network

Xiujuan Lei and Yueyue Wang*

School of Computer Science, Shaanxi Normal University, Xi'an, China

More and more clinical observations have implied that microbes have great effects on human diseases. Understanding the relations between microbes and diseases are of profound significance for disease prevention and therapy. In this paper, we propose a predictive model based on the known microbe-disease associations to discover potential microbe-disease associations through integrating Learning Graph Representations and a modified Scoring mechanism on the Heterogeneous network (called LGRSH). Firstly, the similarity networks for microbe and disease are obtained based on the similarity of Gaussian interaction profile kernel. Then, we construct a heterogeneous network including these two similarity networks and microbe-disease associations' network. After that, the embedding algorithm Node2vec is implemented to learn representations of nodes in the heterogeneous network. Finally, according to these low-dimensional vector representations, we calculate the relevance between each microbe and disease by utilizing a modified rule-based inference method. By comparison with three other methods including LRLSHMDA, KATZHMDA and BiRWHMDA, LGRSH performs better than others. Moreover, in case studies of asthma, Chronic Obstructive Pulmonary Disease and Inflammatory Bowel Disease, there are 8, 8, and 10 out of the top-10 discovered disease-related microbes were validated respectively, demonstrating that LGRSH performs well in predicting potential microbe-disease associations.

Keywords: microbe-disease association, heterogeneous network, network embedding algorithm, Node2vec, skip-gram

OPEN ACCESS

Edited by:
Hyun-Seob Song, University of Nebraska–Lincoln, United States

Reviewed by:
Wen Zhang, Huazhong Agricultural University, China
Sridevi Maharaj, University of California, Irvine, United States

****Correspondence:***
Xiujuan Lei
xjlei@snnu.edu.cn

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *18 November 2019*
Accepted: *17 March 2020*
Published: *15 April 2020*

Citation:
Lei X and Wang Y (2020) Predicting Microbe-Disease Association by Learning Graph Representations and Rule-Based Inference on the Heterogeneous Network. Front. Microbiol. 11:579. doi: 10.3389/fmicb.2020.00579

INTRODUCTION

Varieties of microbial communities are dominant throughout the human different body niches including skin, mouth, respiratory tract, throat, stomach, gut and colon, which mainly compose of bacteria, protozoa, archaeon, viruses, and fungi (Methe et al., 2012; Althani et al., 2016). It is generally that a wide range of them play fundamental roles in human health and diseases such as maintaining homeostasis (Bouskra et al., 2008), developing the immune system (Round and Mazmanian, 2010; Gollwitzer et al., 2014) and resisting pathogens (Methe et al., 2012).

For example, the majority of microbes reside in the gut, regulating human physiology and nutrition by modulating host metabolism and immunity. They can digest and convert dietary constituents into active forms (Qin et al., 2010; Ahn et al., 2013).

Microbial communities are considered as an essential "organ" governing health and disease, which can be influenced by host genetics and host environment such as feeding habits, life styles, seasons and antibiotics (Huttenhower et al., 2012; Althani et al., 2016). If the microbial communities become imbalanced, there may interfere with the symbiotic relationships and cause diseases. For instance, researchers found that the number of phylum Actinobacteria among diabetics was significantly lower than the healthy person (Long et al., 2017). In addition, some studies found a decrease in the relative percentage of Bacteroidetes in obese people compared to the general population (Ley et al., 2006). Moreover, low microbial diversity can lead to inflammatory bowel disease (IBD) (Qin et al., 2010). Thus, understanding the microbe-disease associations can help us know disease pathogenesis to boost disease diagnosis and therapy.

With the advances in sequencing technologies and bioinformatics, more and more microbes living in oceans, soil, human bodies and elsewhere began to be investigated by the scientific community (Gilbert and Dupont, 2011; Methe et al., 2012; Cenit et al., 2014). The Human Microbiome Project Consortium (HMP) was funded to explore the relationships between microbes and human diseases. It generates a wide range of quality-controlled resources and data to develop metagenomic protocols, which is available for scientific research (Methe et al., 2012). Ma et al. (2016) constructed The Human Microbe-Disease Association Database (HMDAD) through collecting correlations between microbes and diseases from 61 published literatures. These achievements provided the foundation for further research on using computational methods to predict potential associations.

In recent years, some computational methods have been conceived for predicting microbe-disease associations based on the assumption that similarly functioning microorganisms incline to share similar associations or non-associations with diseases. By using the Gaussian interaction profile (GIP) kernel similarity, Chen et al. (2017) developed a prediction method called KATZHMDA that infers potential associations based on the number and length of walks in a heterogeneous network. Li et al. (2019) constructed a bidirectional weighted network by combining a normalized Gaussian interaction scheme with a bidirectional recommendation model. Zou et al. (2017) used a bi-random walk and logistic function transformation on a heterogeneous network constructed based on the GIP kernel similarity. Through a combination of the GIP kernel similarity and LapRLS classification, Wang et al. (2017) designed a computing model LRLSHMDA, which is semi-supervised . Meanwhile, through integrating the GIP kernel similarity with disease symptom similarity, Qu et al. (2019) implemented the matrix decomposition and label propagation algorithm on the similarity network for associations' prediction . Huang et al. (2017) predicted potential associations based on known microbe-disease bipartite graph and neighbor collaborative filtering. Moreover, Fan et al. (2019) proposed a method called MDPH_HMDA for prediction by executing standardized HeteSim measurements to weight the relations in a heterogeneous network combined by the GIP kernel similarity, the microbe–microbe functional similarity and the symptom-based human disease similarity. Niu et al. (2019) identified the potential associations by introducing the concept of hypergraph, which put all disease-related microbes on a single hyperedge. In order to take the unequal contributions of microbe and disease information into consider, Zhang et al. (2018) developed a bidirectional similarity integral label propagation method with calculating the microbe functional similarity and the disease semantic similarity.

At the same time, many network embedded methods have been proposed, such as DeepWalk (Perozzi et al., 2014), SDNE (Wang et al., 2016), Node2vec (Grover and Leskovec, 2016), etc. In this study, inspired by the performance of graph representations for many real-world problems such as protein network research, text and visual processing (Cao et al., 2016). We utilize Node2vec (Grover and Leskovec, 2016) to predict potentially unknown associations (LGRSH) on a heterogeneous network. First, similarity networks for microbes and diseases are calculated by the GIP kernel similarity. Then, we construct a heterogeneous network integrating the two similarity networks and known microbe-disease associations' network. After that, the embedding algorithm Node2vec has been utilized to assign a low-dimensional vector representation to nodes in the heterogeneous network. Finally, according to the vector representation of each node, we calculate the degrees of correlation between microbes and diseases to discover potential associations with a modified rule-based inference method. In order to assess the prediction performance of LGRSH, we implemented Leave-one-out cross validation (LOOCV) and fivefold cross validation. The area under the receiver operating characteristic curve (AUC) obtained by LGRSH are 0.9260 and 0.9254, which is better than the compared methods. Moreover, case studies of asthma, Chronic Obstructive Pulmonary Disease (COPD) and IBD demonstrate that LGRSH can be considered as an effective method for association prediction.

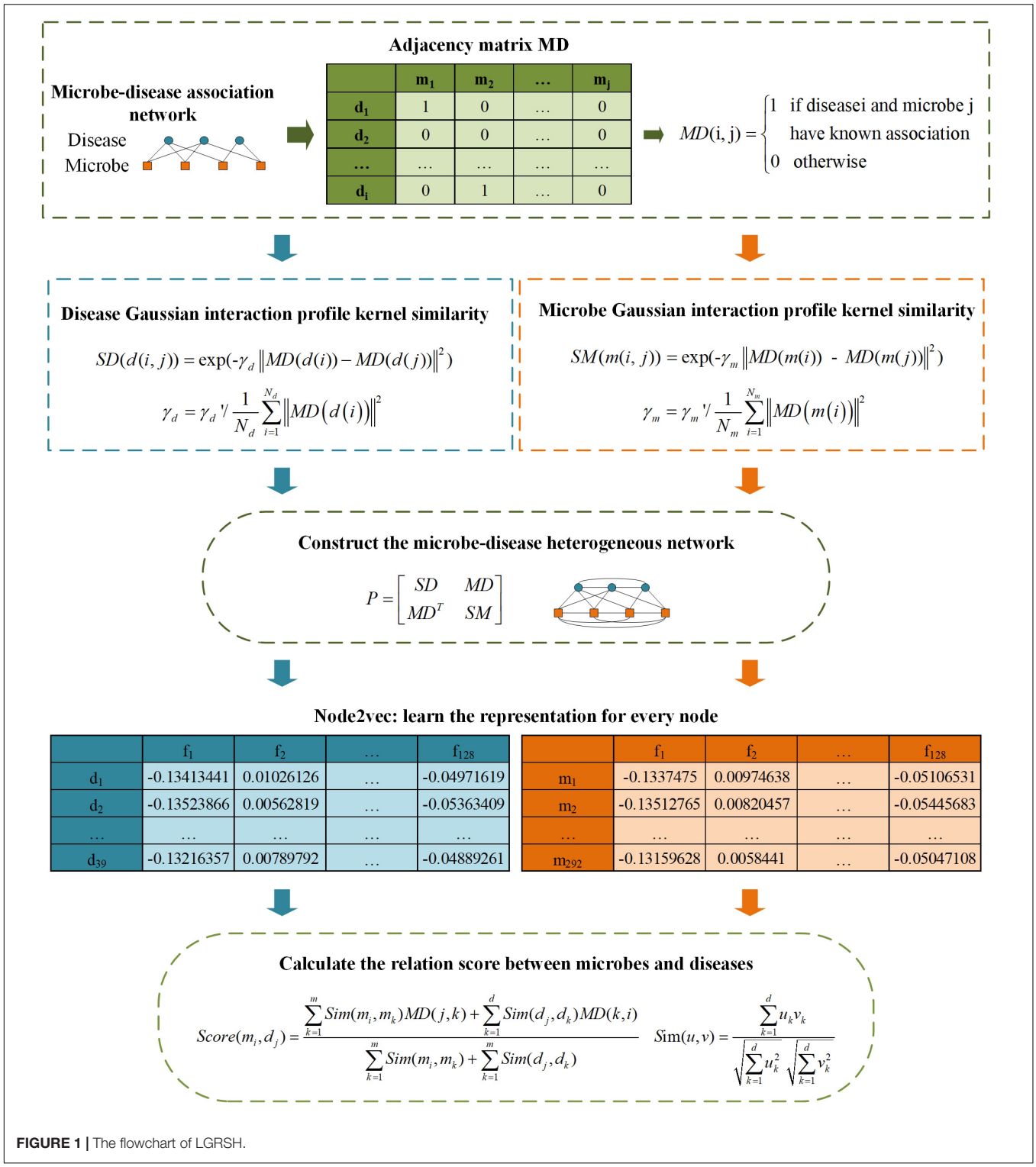
MATERIALS AND METHODS

Material

We download microbe-disease associations from HMDAD (Ma et al., 2016), which contains 483 verified associations' records between 292 microbes and 39 diseases. After removing the repetitive relationships, 450 distinct associations' records are obtained. Then we construct a 39×292 dimensional adjacency matrix MD of the associations' network. $MD(i, j)$ is 1 indicating that there is a known association between disease $d(i)$ and microbe $m(j)$, otherwise, $MD(i, j)$ is 0.

Methods

As illustrated in **Figure 1**, firstly, the similarity networks for microbe and disease have been constructed. And then, a heterogeneous network integrating two similarity networks and



	m_1	m_2	…	m_j
d_1	1	0	…	0
d_2	0	0	…	0
…	…	…	…	…
d_i	0	1	…	0

	f_1	f_2	…	f_{128}
d_1	-0.13413441	0.01026126	…	-0.04971619
d_2	-0.13523866	0.00562819	…	-0.05363409
…	…	…	…	…
d_{39}	-0.13216357	0.00789792	…	-0.04889261

	f_1	f_2	…	f_{128}
m_1	-0.1337475	0.00974638	…	-0.05106531
m_2	-0.13512765	0.00820457	…	-0.05445683
…	…	…	…	…
m_{292}	-0.13159628	0.0058441	…	-0.05047108

FIGURE 1 | The flowchart of LGRSH.

microbe-disease associations' network can be obtained. After that, the embedding algorithm Node2vec is utilized to learn the representation for every node. Finally, according to the topology information based on Node2vec method, we calculate the relation score between every microbe vector and disease vector.

Calculation of Microbe Similarities Based on the GIP Kernel Similarity

Based on the assumption that two microbes are more likely to share functional similarities potentially if they are related to more common diseases. We calculate the GIP kernel similarity for

microbes based on known microbe-disease associations' network. For microbes *m(i)* and *m(j)*, the similarity score is obtained according to Eq. (1) (Wang et al., 2017):

$$SM(m(i,\ j))\ =\ \exp(-\gamma_m\|MD(m(i))\ -\ MD(m(j))\|^2) \quad (1)$$

where *m(i, j)* represents two arbitrary microbes in matrix *MD*. Parameter γ_m is used to control the bandwidth and is affected by a new bandwidth parameter γ_m'(Wang et al., 2017), which can be obtained as Eq. (2):

$$\gamma_m = \gamma'_m / \frac{1}{N_m}\sum_{i=1}^{N_m}\|MD\,(m\,(i))\,\|^2 \quad (2)$$

here, N_m is equal to 292, which indicates the total number of microbes. The parameter γ_m' is set to 1 for simplicity (Wang et al., 2017).

Calculation of Disease Similarities Based on the GIP Kernel Similarity

In the similar way, we construct a disease similarity network by using the GIP kernel similarity for each disease pair. The similarity between disease *d(i)* and *d(j)* is obtained according to Eq. (3) (Wang et al., 2017):

$$SD(d(i,j))=\ \exp(-\gamma_d\|MD(d(i))-MD(d(j))\|^2) \quad (3)$$

where *d(i, j)* represents two arbitrary diseases in matrix *MD*. The parameter γ_d can be obtained as Eq. (4):

$$\gamma_d = \gamma'_d / \frac{1}{N_d}\sum_{i=1}^{N_d}\|MD\left(d\,(i)\right)\|^2 \quad (4)$$

here, N_d is equal to 39, which indicates the total number of diseases. The parameter γ_d' is set to 1 for simplicity (Wang et al., 2017).

Constructing a Heterogeneous Network for Microbes and Diseases

According to the Eqs (1) and (3), we have constructed two similarity matrices SM and SD. Then we construct a heterogeneous network including the edges of microbe–microbe, microbe-disease and disease–disease associations, and it can be expressed as Eq. (5):

$$P=\begin{bmatrix} SD & MD \\ MD^T & SM \end{bmatrix} \quad (5)$$

where *P* represents the matrix of heterogeneous network. MD^T is the transpose of *MD*.

Using Node2vec to Learning Representations

Node2vec is a flexible neighborhood sampling strategy which can explore neighborhoods in the form of Breadth-First Sampling (*BFS*) and Depth-First Sampling (*DFS*) fashion by introducing two parameters (Grover and Leskovec, 2016). It maximizes the network neighborhood of nodes by mapping nodes to vector feature spaces. Therefore, we apply Node2vec to learn vector representations for nodes in the heterogeneous network.

Firstly, we utilize a bias random walk strategy to calculate the transition probabilities for every node. For a current node *u*, the probability of accessing the next node *x* can be calculated as follows:

$$P(c_i=x|c_{i-1}=u)=\begin{cases}\frac{\pi_{ux}}{Z} & if\ (u,x)\in E\\ 0 & otherwise\end{cases} \quad (6)$$

here, *Z* is a regularization constant. π_{ux} is denormalized transition probabilities on edges (*u*, *x*) leading from *u*, which is influenced by a weight adjustment parameter α. We suppose the walk just went from *t* to *u* and set$\pi_{ux}=\alpha_{pq}\,(t,\ x)\cdot w_{ux}$, where

$$a_{pq}(t,\ x)=\begin{cases}1/p & if\ d_{tx}=0\\ 1 & if\ d_{tx}=1\\ 1/q & if\ d_{tx}=2\end{cases} \quad (7)$$

here, d_{tx} is in the range of {0, 1, 2}, representing the shortest distance from nodes *t* to *x*. Parameters *p* and *q* are used to strike a balance between *DFS* and *BFS*. As shown in **Figure 2**, parameter *p* is a return parameter that affects the possibility of re-traversing a node immediately during a walk. If *p* is set to be larger, it is less likely to revisit the node that was just accessed. This strategy can lead to moderate exploration and avoid repetitive sampling. If the value is set to be smaller, the walk is more likely to backtrack, and tends to reach nodes near the node. There is more concerned for the local information. Parameter *q* is an in-out parameter, which allows searches to distinguish "inward" and "outward" nodes (Zeng et al., 2019). If $q > 1$, the walk tends to be closer to node *u*. In contrast, if $q < 1$, it tends to traverse nodes far from node *u* (Zeng et al., 2019).

We first select one node *u* and mark it as the current node, and then select one node *v* from all the neighbors of the current node *u* based on the transition probabilities calculated above. Following, we mark this newly selected node *v* as the current node and repetitive such as a node sampling process. The algorithm terminates when the number of nodes in a sequence reaches a preset walking length *l*. By referring to the previous paper, we set *l* as10 (Munui et al., 2018).

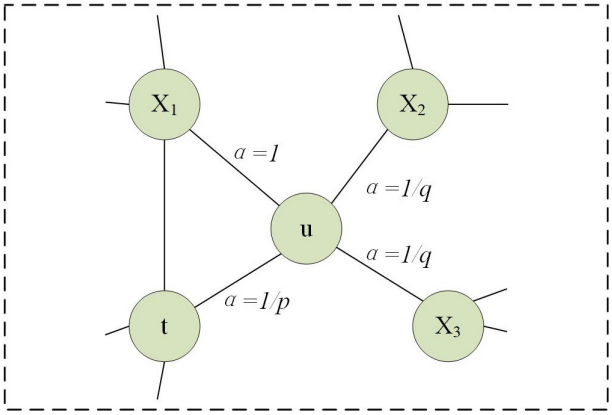


FIGURE 2 | Description of walking strategy in Node2vec when the traversal has just gone from *t* to *u*.

Node2vec uses Skip-gram model to generate eigenvectors of nodes (Jang et al., 2019). Skip-gram model is a word embedding algorithms for learning distributed vector representations from a large number of textual corpora which tries to categorize a word according to other words in the same sentence as much as possible (Mikolov et al., 2013). In fact, the sequence of nodes obtained by bias random walk algorithm, each node actually corresponds to a word. The input of this model is the sequence encoding of a node, and the output is the nodes before and after the sequence. In this paper, we set the context size to 10 and the dimension of these eigenvectors to 128 according to the original parameter selection for the best performance (Grover and Leskovec, 2016). The algorithm is detailed in **Figure 3**.

Association Discovering

According to the popular rule-based inference method for predicting novel drug-target associations based on indirect relationships in 2017 (Zong et al., 2017), we utilize a modified Scoring mechanism to grade microbe-disease relations based on the low-dimensional vector representation. Considering that indirect relationships do not fully predict the relationship if there are few known relations between some microbes and diseases, especially if there is only single relationship, we have used both direct and indirect connections to calculate correlations between microbes and diseases.

We use $Score(m_i, d_j)$ to represent the correlation score between the *ith* microbe and *jth* disease in the heterogeneous network. It can be calculated according to Eq. (8):

$$Score(m_i, d_j) = \frac{\sum_{k=1}^{m} Sim(m_i, m_k)MD(j, k) + \sum_{k=1}^{d} Sim(d_j, d_k)MD(k, i)}{\sum_{k=1}^{m} Sim(m_i, m_k) + \sum_{k=1}^{m} Sim(d_j, d_k)} \quad (8)$$

In this Equation, m and d indicate the numbers of microbe and disease, $MD(i, j)$ is the association between disease i and microbe j. The $Sim(u, v)$ is calculated as Eq. (9):

$$Sim(u, v) = \frac{\sum_{k=1}^{d} u_k v_k}{\sqrt{\sum_{k=1}^{d} u_k^2} \sqrt{\sum_{k=1}^{d} v_k^2}} \quad (9)$$

here, d represents the dimension for each vector, u_k, v_k represent the components of vectors u and v.

Learning representations
Input: Graph $P = (V, E, W)$, dimension d, walks per node r, walk length l, Context size k, Return parameter p, In-out parameter q
Output: Eigenvectors of each node

```
Node2vec(P, d, r, l, k, p, q)
    π = TP preprocessing (P, p, q)          // Calculate the transition probabilities
    P' = (V, E, π)                          // Normalize transition probabilities
    Initialize walks to Empty
    For iter = 1 to r:
        for all nodes u ∈ V:                // Simulate a random walk starting from start node
            Initialize walk to [u]
            for walk_length = 1 to l:
                curr = walk[- 1]
                Vcurr = Get Neighbors(curr, P')   // Find neighbors of the current node
                s = probability select(Vcurr, π)  // Walk based on transition probabilities
                Append s to walk
    Append walk to walks
    F = Skip-gram (k, d, walks)             // Generate eigenvectors of each node
    return F
```

FIGURE 3 | Description of algorithm Node2vec.

TABLE 1 | Effect of parameters p and q in fivefold cross validation.

	q = 0.25	q = 0.5	q = 1	q = 2	q = 4	q = 8	q = 16
p = 0.25	0.9251	0.9165	0.9178	0.9246	0.9229	0.9236	0.9244
p = 0.5	0.9253	0.9236	0.9251	0.9246	**0.9254**	0.9235	0.9229
p = 1	0.9240	0.9250	0.9190	0.9213	0.9234	0.9234	0.9242
p = 2	0.9214	0.9204	0.9239	0.9230	0.9251	0.9181	0.9208
p = 4	0.9215	0.9222	0.9206	0.9229	0.9241	0.9239	0.9235

Bold values: LGRSH achieves the best performance while p = 0.5, q = 4.

RESULTS

We implement LOOCV and fivefold cross validation on HMDAD to assess the prediction performance of LGRSH. In the LOOCV, we regard each known association as a test sample, with other known associations as training samples (Quan et al., 2014). All unverified microbe-disease associations are regarded as candidate samples. In the fivefold cross validation, we randomly divide all known microbe-disease associations into 5 average groups. Each of these five groups is regarded as testing sample, while other four groups are training samples. This process is conducted five times to mitigate the bias due to random sample partitioning (Niu et al., 2019). Based on the prediction score, we evaluate the predictive performance by ranking the test samples. The AUC can be calculated according to the receiver operating characteristic (ROC) curve. If there is a random prediction performance, the AUC value is 0.5.

Effect of Parameters

There are two important parameters in Node2vec. One is a return parameter p and another is an in-out parameter q. We set various values under the framework of fivefold cross validation in order to evaluate the impact of these parameters. According to the comparison results in **Table 1** and **Figure 4**, we can find that the performance of LGRSH is best with 0.9254 while $p = 0.5$, $q = 4$. Hence, we set $p = 0.5$, $q = 4$ in the subsequent experiments.

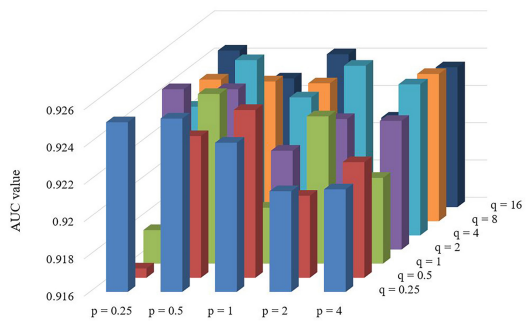


FIGURE 4 | Effect of parameters p and q in fivefold cross validation.

Comparison With Other Methods

We compare LGRSH with three methods including LRLSHMDA (Wang et al., 2017), KATZHMDA (Chen et al., 2017) and BiRWHMDA (Zou et al., 2017). These four methods are measured by Precision-recall curve. As illustrated in **Figures 5**, **6**, we can draw a conclusion that LGRSH performs better than other three methods.

Furthermore, we measure the top-level results of LGRSH and three other methods in LOOCV. As shown in **Figure 7**, LGRSH can find more known associations among the top 500 predicted microbes.

CASE STUDIES

To evaluate the ability of LGRSH for discovering unknown associations in HMDAD, we implement case studies in asthma, COPD and IBD. We conduct experiments for 10 times on each diseases to make the results more stable. After calculating the similarity of every microbe and disease, the scores are sorted in descending order to obtain the top-10 candidate microbes for every disease. The scores of top-10 disease-related microbes are provided in **Supplementary Tables S1–S3**, respectively.

Asthma

Asthma is a common inflammatory disease affecting more than 300 million people all over the world, which is more common in childhood with recurrent cough, wheezing and breathing difficulties. In recent years, asthma has been found to be closely linked with microbes (Caliskan et al., 2013). Hence, we consider Asthma for case studies. As shown in **Table 2**, 8 of top-10 discovered microbes were confirmed. For instance, Clostridium difficile colonization (ranked 1st in the list) in 1 month was associated with asthma between the ages of 6 and

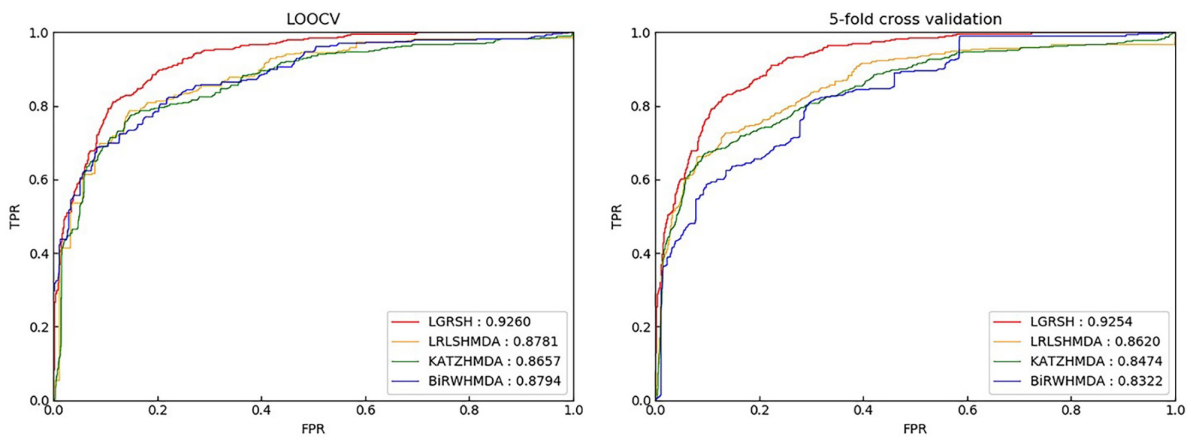


FIGURE 5 | Prediction comparison between LGRSH and other three methods in LOOCV and fivefold cross validation while $p = 0.5$, $q = 4$.

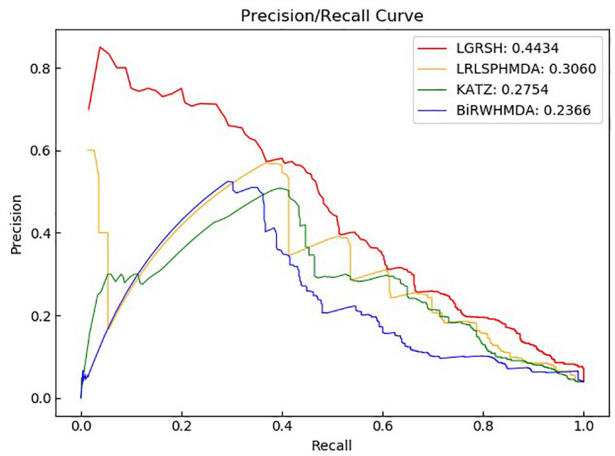


FIGURE 6 | Precision-recall curves for LGRSH and other three methods in fivefold cross validation.

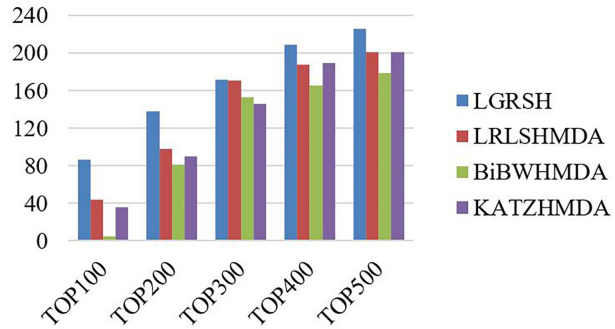


FIGURE 7 | The number of correctly predicted by LGRSH and other three methods on HMDAD.

TABLE 2 | Validation results for Top-10 predicted microbes related with asthma.

Rank	Microbe	Evidence
1	Clostridium difficile	PMID:21872915
2	Firmicutes	PMID:27078029
3	Clostridium coccoides	PMID:21477358
4	Actinobacteria	PMID:30286807
5	Enterobacteriaceae	PMID:28947029
6	Lactobacillus	PMID:30400588
7	Bacteroides	PMID:18822123
		PMID:29161087
8	Burkholderia	**Unconfirmed**
9	Lachnospiraceae	PMID:28912020
10	Enterococcus	**Unconfirmed**

7 (van Nimwegen et al., 2011). Researchers also proved that colonization with Clostridium coccoides (ranked 3rd in the list) and Bacteroides (ranked 7th in the list) at 3 weeks were associated with positive predictors of asthma at age 3 (Carl et al., 2008, 2011). In addition, the abundance of Firmicutes (ranked 2nd in the list) and Enterobacteriaceae (ranked 5th in the list) were higher in severe asthmatics compared with non-asthmatic people, while Actinobacteria (ranked 4th in the list) and Lachnospiraceae (ranked 9th in the list) with lower proportion (Marri et al., 2013; Ciaccio et al., 2015; Zhang et al., 2016; Li et al., 2017). Moreover, Huang et al. (2018) found that Lactobacillus (ranked 6th in the list) can reduce asthma severity and improve asthma control, which is beneficial to children with asthma.

TABLE 3 | Validation results for Top-10 predicted microbes related with COPD.

Rank	Microbe	Evidence
1	Proteobacteria	PMID:29579057
2	Prevotella	PMID:28542929
3	Helicobacter pylori	PMID:28558695
4	Actinobacteria	PMID:29709671
5	Bacteroidetes	PMID:29579057
6	Clostridium difficile	PMID:30430993
7	Clostridium coccoides	**Unconfirmed**
8	Lactobacillus	PMID:26630356
9	Lachnospiraceae	**Unconfirmed**
10	Staphylococcus aureus	PMID:30804927

TABLE 4 | Validation results for Top-10 predicted microbes related with IBD.

Rank	Microbe	Evidence
1	Prevotella	PMID:24013298
2	Bacteroidetes	PMID:29492876
3	Clostridium difficile	PMID:24838421
4	Helicobacter pylori	PMID:22221289
		PMID:28124160
5	Firmicutes	PMID:25307765
		PMID:29492876
6	Clostridium coccoides	PMID:19235886
7	Lactobacillus	PMID:26340825
8	Enterobacteriaceae	PMID:30319571
9	Veillonella	PMID:30573380
10	Haemophilus	PMID:24013298

Chronic obstructive pulmonary disease (COPD)

Chronic obstructive pulmonary disease is a progressive obstructive pulmonary disease with main symptoms of breathing difficultly and coughing (Rabe et al., 2007). It is more common among smokers, and is also influenced by factors like air pollution and genetics. Although the disease can be slowed down by treatment, there is still no clear treatment or pathogenesis for it. Recently, some findings indicate that changes in microbes may have significant effects in the development of COPD (Malhotra and Henric, 2015). Thus, we consider COPD for case studies. As shown in **Table 3**, 8 of top 10 discovered microbes were confirmed. For example, the main flora of Proteobacteria (ranked 1st in the list) and Bacteroidetes (ranked 5th in the list) increased with the deterioration of COPD (Rohde et al., 2004). Researchers also found that Helicobacter pylori (ranked 3rd in the list)

infection is associated with reduced lung function and systemic inflammation in COPD patients (Mammen and Sethi, 2016). In patients with COPD, the proportion of Prevotella (ranked 2nd in the list) is reduced compared with healthy people, but phyla Actinobacteria (ranked 4th in the list), Clostridium difficile (ranked 6th in the list) and Lactobacillus (ranked 8th in the list) are increased (Yadava et al., 2016; Larsen, 2017; de Miguel-Diez et al., 2018; Ghebre et al., 2018). For example, the Clostridium difficile is twice as high in COPD patients as in healthy person. Moreover, Staphylococcus aureus (ranked 10th in the list) has been found in the respiratory tract of patients with COPD (Uddin et al., 2019).

Inflammatory bowel disease (IBD)

Inflammatory bowel disease is a chronic, idiopathic gastrointestinal inflammatory disease that is thought to be influenced by environmental and host factors (D'Aoust et al., 2017). It is characterized by recurrent episodes, diverse clinical manifestations and severe complications such as bleeding, abscess formation and perforation (Cosnes et al., 2002). In this paper, we consider IBD for case studies. As shown in **Table 4**, 10 of top-10 discovered microbes were confirmed. For instance, researchers have found that IBD is related to gut microbiological disorders including expansion of Enterobacteriaceae facultative anaerobic bacteria (ranked 8th in the list) and decrease in some beneficial fecal bacteria such as Firmicutes (ranked 5th in the list) (Eom et al., 2018; Zuo and Ng, 2018). In patients with IBD, the dominant of Prevotella (ranked 1st in the list), Veillonella (ranked 9th in the list) and Haemophilus (ranked 10th in the list) were largely contribute to dysbiosis (Said et al., 2014). Bacteroidetes (ranked second in the list) and Lactobacillus (ranked 7th in the list) were significantly increased compared with healthy people, but the Clostridium coccoides (ranked 6th in the list) was less abundant (Sokol et al., 2009; Thomas et al., 2015; Eom et al., 2018). Researchers also found that Clostridium difficile (ranked 3rd in the list) infection has become a significant clinical challenge for patients suffering from IBD, which can worsen flares of IBD, inducing to emergent colectomies and mortality (Hashash and Binion, 2014). Moreover, recent experimental results found that chronic infection with Helicobacter pylori (ranked 4th in the list) is protective against IBD. And IBD patients are least likely to be infected with Helicobacter pylori compared to the normal population (Sonnenberg and Genta, 2012; Kyburz and Muller, 2017).

CONCLUSION

There are countless microbe communities inhabited in the human body, having important impacts on human health and disease by regulating the metabolism and immunity. With the establishment of relational databases for microbes and diseases, exploring their associations have become a hot topic for researchers. In this study, we propose a predictive approach called LGRSH by utilizing network embedding algorithm Node2vec to obtain the representation for every node in the heterogeneous network. According to the vector representation for every node, we rank the relevance of each microbe vector and disease vector to discover potential microbe-disease associations. In LOOCV and 5-fold cross validation, LGRSH performs better compared with three other methods with AUC reached 0.9260 and 0.9254. The case studies of asthma, COPD and IBD show that LGRSH can be used as a predictive tool for microbe-disease associations.

Certainly, there are still some deficiencies in LGRSH. For example, there are only 450 know micro-disease associations, which accounts for very small proportion of human microbial diseases. This may result in less comprehensive for prediction. We believe that the problem will be solved when more microbe-disease links are discovered. In addition, the embedding algorithm itself is a local method. In the future, we will learn more graph representation algorithms to improve the global capability. Moreover, we calculate the similarities for microbe and disease through the GIP kernel, which may biased toward microbes and diseases with more known associations. Hence, we will improve the efficiency of LGRSH by integrating some optimization strategies such as microbe functional similarity, disease semantic similarity and symptom-based disease similarity in the future work.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

XL and YW conceptualized the study and read and approved the final manuscript. YW conducted the experiments, analyzed the result, and wrote the manuscript. XL conceived the project, analyzed the result, and revised the manuscript.

FUNDING

This work was supported by the funding from National Natural Science Foundation of China (Nos. 61972451, 61672334, and 61902230) and the Fundamental Research Funds for the Central Universities (No. GK201901010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.00579/full#supplementary-material

REFERENCES

Ahn, J., Sinha, R., Pei, Z. H., Dominianni, C., Wu, J., Shi, J. X., et al. (2013). Human gut microbiome and risk for colorectal Cancer. *J. Natl. Cancer Inst.* 105, 1907–1911. doi: 10.1093/jnci/djt300

Althani, A. A., Marei, H. E., Hamdi, W. S., Nasrallah, G. K., El Zowalaty, M. E., Al Khodor, S., et al. (2016). Human microbiome and its association with health and diseases. *J. Cell. Physiol.* 231, 1688–1694. doi: 10.1002/jcp.25284

Bouskra, D., Brezillon, C., Berard, M., Werts, C., Varona, R., Boneca, I. G., et al. (2008). Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. *Nature* 456, 507–534. doi: 10.1038/nature07450

Caliskan, M., Bochkov, Y. A., Kreiner-Moller, E., Bonnelykke, K., Ober, C., et al. (2013). Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N. Engl. J. Med.* 368, 1398–1407. doi: 10.1056/NEJMoa1211592

Cao, S., Lu, W., and Xu, Q. (2016). *Deep Neural Networks for Learning Graph Representations. Paper presented at the AAAI*. Menlo Park, CA: AAAI Press.

Carl, V., Liesbeth, V., Kristine, N. D., and Herman, G. (2011). Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol.* 11:68. doi: 10.1186/1471-180-11-68

Carl, V., Vera, N., Verhulst, S. L., Herman, G., and Medicine, Desager, K. N. (2008). Early intestinal *Bacteroides* fragilis colonisation and development of asthma. *BMC Pulmon. Med.* 8:19. doi: 10.1186/1471-2466-8-19

Cenit, M. C., Matzaraki, V., Tigchelaar, E. F., and Zhernakova, A. (2014). Rapidly expanding knowledge on the role of the gut microbiome in health and disease. *Biochim. Biophy. Acta Mol. Basis Dis.* 1842, 1981–1992. doi: 10.1016/j.bbadis.2014.05.023

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2017). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33, 733–739. doi: 10.1093/bioinformatics/btw715

Ciaccio, C. E., Barnes, C., Kennedy, K., Chan, M., Portnoy, J., and Rosenwasser, L. (2015). Home dust microbiota is disordered in homes of low-income asthmatic children. *J. Asthma* 52, 1–8. doi: 10.3109/02770903.2015.1028076

Cosnes, J., Cattan, S., Blain, A., Beaugerie, L., Carbonnel, F., Parc, R., et al. (2002). Long-term evolution of disease behavior of Crohn's disease. *Inflamm. Bowel Dis.* 8, 244–250. doi: 10.1097/00054725-200207000-00002

D'Aoust, J., Battat, R., and Bessissow, T. (2017). Management of inflammatory bowel disease with clostridium difficile infection. *World J. Gastroenterol.* 23, 4986–5003. doi: 10.3748/wjg.v23.i27.4986

de Miguel-Diez, J., Lopez-de-Andres, A., Esteban-Vasallo, M. D., Hernandez-Barrera, V., de Miguel-Yanes, J. M., Mendez-Bailon, M., et al. (2018). Clostridium difficile infection in hospitalized patients with COPD in Spain (2001-2015). *Eur. J. Intern. Med.* 57, 76–82. doi: 10.1016/j.ejim.2018.06.022

Eom, T., Kim, Y. S., Choi, C. H., Sadowsky, M. J., and Unno, T. (2018). Current understanding of microbiota- and dietary-therapies for treating inflammatory bowel disease. *J. Microbiol.* 56, 189–198. doi: 10.1007/s12275-018-8049-8

Fan, C. Y., Lei, X. J., Guo, L., and Zhang, A. D. (2019). Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. *Neurocomputing* 323, 76–85. doi: 10.1016/j.neucom.2018.09.054

Ghebre, M. A., Pang, P. H., Diver, S., Desai, D., Bafadhel, M., Haldar, K., et al. (2018). Biological exacerbation clusters demonstrate asthma and chronic obstructive pulmonary disease overlap with distinct mediator and microbiome profiles. *J. Allergy Clin. Immunol,* 141, 2027.e12–2036.e12. doi: 10.1016/j.jaci.2018.04.013

Gilbert, J. A., and Dupont, C. L. (2011). Microbial metagenomics: beyond the genome. *Annu. Rev. Mar. Sci.* 3, 347–371. doi: 10.1146/annurev-marine-120709-142811

Gollwitzer, E. S., Saglani, S., Trompette, A., Yadava, K., Sherburn, R., McCoy, K. D., et al. (2014). Lung microbiota promotes tolerance to allergens in neonates via PD-L1. *Nat. Med.* 20, 642–647. doi: 10.1038/nm.3568

Grover, A., and Leskovec, J. (2016). node2vec: scalable feature learning for networks. *KDD* 2016, 855–864. doi: 10.1145/2939672.2939754

Hashash, J. G., and Binion, D. G. (2014). Managing clostridium difficile in inflammatory bowel disease (IBD). *Curr. Gastroenterol. Rep.* 16:393. doi: 10.1007/s11894-014-0393-1

Huang, C.-F., Chie, W.-C., and Wang, I.-J. J. N. (2018). Efficacy of lactobacillus administration in school-age children with asthma: a randomized, placebo-controlled trial. *Nutrients* 10:1678. doi: 10.3390/nu10111678

Huang, Y. A., You, Z. H., Chen, X., Huang, Z. A., Zhang, S., and Yan, G. Y. (2017). Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* 15:209. doi: 10.1186/s12967-017-1304-7

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Jang, B., Kim, I., and Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *Plos One* 14:e0220976. doi: 10.1371/journal.pone.0220976

Kyburz, A., and Muller, A. (2017). *Helicobacter* pylori and extragastric diseases. *Curr. Top. Microbiol. Immunol.* 400, 325–347. doi: 10.1007/978-3-319-50520-6_14

Larsen, J. M. (2017). The immune response to *Prevotella bacteria* in chronic inflammatory disease. *Immunology* 151, 363–374. doi: 10.1111/imm.12760

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023. doi: 10.1038/4441022a

Li, H., Wang, Y. Q., Jiang, J. W., Zhao, H. C., Feng, X., Wang, L., et al. (2019). A novel human microbe-disease association prediction method based on the bidirectional weighted network. *Front. Microbiol.* 10:676. doi: 10.3389/fmicb.2019.00676

Li, N., Qiu, R., Yang, Z., Li, J., Chung, K. F., Zhang, Q. J. R. M., et al. (2017). Sputum microbiota in severe asthma patients: relationship to eosinophilic inflammation. *Respiratory Med.* 131, 192–198. doi: 10.1016/j.rmed.2017.08.016

Long, J., Cai, Q., Steinwandel, M., Hargreaves, M. K., Bordenstein, S. R., Blot, W. J., et al. (2017). Association of oral microbiome with type 2 diabetes risk. *J. Periodontal Res.* 52, 636–643. doi: 10.1111/jre.12432

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., and Cui, Q. (2016). An analysis of human microbe–disease associations. *Briefings Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Malhotra, R., and Henric, O. (2015). Immunology, genetics and microbiota in the COPD pathophysiology: potential scope for patient stratification. *Expert Rev. Respirat. Med.* 9, 153–159. doi: 10.1586/17476348.2015.1000865

Mammen, M. J., and Sethi, S. (2016). COPD and the microbiome. *Respirology* 21, 590–599. doi: 10.1111/resp.12732

Marri, P. R., Stern, D. A., Wright, A. L., Billheimer, D., and Martinez, F. D. (2013). Asthma-associated differences in microbial composition of induced sputum. *J. Allergy Clin. Immunol.* 131, 346.–352.e1-13. doi: 10.1016/j.jaci.2012.11.013

Methe, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., et al. (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Science, J. D. J. C. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. Ithaca, NY: Cornell University.

Munui, K., Han, B. S., and Min, S. (2018). Relation extraction for biological pathway construction using node2vec. *Bmc Bioinform.* 19:206. doi: 10.1186/s12859-018-2200-8

Niu, Y. W., Qu, C. Q., Wang, G. H., and Yan, G. Y. (2019). RWHMDA: random walk on hypergraph for microbe-disease association prediction. *Front. Microbiol.* 10:1578. doi: 10.3389/fmicb.2019.01578

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: online learning of social representations," in *Paper presented at the Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and data Mining*, New York, NY: ACM.

Qin, J., Li, R., Raes, J., Arumugam, M., and Nature, M. K. J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Qu, J., Zhao, Y., and Yin, J. (2019). Identification and analysis of human microbe-disease associations by matrix decomposition and label propagation. *Front. Microbiol.* 10:291. doi: 10.3389/fmicb.2019.00291

Quan, Z., Jinjin, L., and Chunyu, W. (2014). Approaches for Recognizing Disease Genes Based on Network. *Biomed. Res. Int.* 2014, 416323. doi: 10.1155/2014/416323

Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., and Respiratory, J. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD Executive Summary. *Am. J. Respir. Crit. Care Med.* 176, 532–555. doi: 10.1164/rccm.200703-456SO

Rohde, G., Gevaert, P., Holtappels, G., Borg, I., Wiethege, A., Arinir, U., et al. (2004). Increased IgE-antibodies to Staphylococcus aureus enterotoxins in patients with COPD. *Respir. Med.* 98, 858–864. doi: 10.1016/j.rmed.2004.02.012

Round, J. L., and Mazmanian, S. K. (2010). Inducible Foxp3+ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12204–12209. doi: 10.1073/pnas.0909122107

Said, H. S., Suda, W., Nakagome, S., Chinen, H., Oshima, K., Kim, S., et al. (2014). Dysbiosis of salivary microbiota in inflammatory bowel disease and its association with oral immunological biomarkers. *DNA Res.* 21, 15–25. doi: 10.1093/dnares/dst037

Sokol, H., Seksik, P., Furet, J. P., Firmesse, O., Nion-Larmurier, I., Beaugerie, L., et al. (2009). Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflamm. Bowel. Dis.* 15, 1183–1189. doi: 10.1002/ibd.20903

Sonnenberg, A., and Genta, R. M. (2012). Low prevalence of *Helicobacter* pylori infection among patients with inflammatory bowel disease. *Aliment. Pharmacol. Ther.* 35, 469–476. doi: 10.1111/j.1365-2036.2011.04969.x

Thomas, M., Langella, P., and Neyrolles, O. (2015). Lactobacillus acidophilus: a promising tool for the treatment of inflammatory bowel diseases?. *Med. Sci.* 31, 715–717. doi: 10.1051/medsci/20153108004

Uddin, M., Watz, H., Malmgren, A., and Pedersen, F. (2019). NETopathic inflammation in chronic obstructive pulmonary disease and severe asthma. *Front. Immunol.* 10:47. doi: 10.3389/fimmu.2019.00047

van Nimwegen, F. A., Penders, J., Stobberingh, E. E., Postma, D. S., Koppelman, G. H., Kerkhof, M., et al. (2011). Mode and place of delivery, gastrointestinal microbiota, and their influence on asthma and atopy. *J. Allergy Clin. Immunol.* 128, 948–955.e1-e3. doi: 10.1016/j.jaci.2011.07.027

Wang, D., Cui, P., and Zhu, W. (2016). "Structural deep network embedding," in *Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM.

Wang, F., Huang, Z. A., Chen, X., Zhu, Z. X., Wen, Z. K., Yan, G. Y., et al. (2017). LRLSHMDA: laplacian regularized least squares for human microbe-disease association prediction. *Sci. Rep.* 7:7601. doi: 10.1038/s41598-017-08127-2

Yadava, K., Pattaroni, C., Sichelstiel, A. K., Trompette, A., Gollwitzer, E. S., Salami, O., et al. (2016). Microbiota promotes chronic pulmonary inflammation by enhancing il-17a and autoantibodies. *Am. J. Respir. Crit. Care Med.* 193, 975–987. doi: 10.1164/rccm.201504-0779OC

Zeng, M., Li, M., Fei, Z., Wu, F., Li, Y., Pan, Y., et al. (2019). A deep learning framework for identifying essential proteins by integrating multiple types of biological information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* PP:1. doi: 10.1109/TCBB.2019.2897679

Zhang, Q., Cox, M., Liang, Z., Brinkmann, F., Cardenas, P. A., Duff, R., et al. (2016). Airway microbiota in severe asthma and relationship to asthma severity and phenotypes. *PloS One* 11:e0152724. doi: 10.1371/journal.pone.0152724

Zhang, W., Yang, W. T., Lu, X. T., Huang, F., and Luo, F. (2018). the bi-direction similarity integration method for predicting microbe-disease associations. *IEEE Access.* 6, 38052–38061. doi: 10.1109/ACCESS.2018.2851751

Zong, N., Kim, H., Ngo, V., and Harismendy, O. J. B. (2017). Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics* 33, 2337–2344. doi: 10.1093/bioinformatics/btx160

Zou, S., Zhang, J. P., and Zhang, Z. P. (2017). A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. *PloS One* 12:e0184394. doi: 10.1371/journal.pone.0184394

Zuo, T., and Ng, S. C. (2018). The gut microbiota in the pathogenesis and therapeutics of inflammatory bowel disease. *Front. Microbiol.* 9:2247. doi: 10.3389/fmicb.2018.02247

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

METHODS
published: 06 May 2020
doi: 10.3389/fmicb.2020.00648

Assessing and Interpreting the Metagenome Heterogeneity With Power Law

Zhanshan (Sam) Ma[1,2]*

[1] Computational Biology and Medical Ecology Lab, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China, [2] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

There are two major sequencing technologies for investigating the microbiome: the amplicon sequencing that generates the OTU (operational taxonomic unit) tables of marker genes (e.g., bacterial 16S-rRNA), and the metagenomic shotgun sequencing that generates metagenomic gene abundance (MGA) tables. The OTU table is the counterpart of species abundance tables in macrobial ecology of plants and animals, and has been the target of numerous ecological and network analyses in recent gold rush for microbiome research and in great efforts for establishing an inclusive theoretical ecology. Nevertheless, MGA analyses have been largely limited to bioinformatics pipelines and *ad hoc* statistical methods, and systematic approaches to MGAs guided by classic ecological theories are still few. Here, we argue that, the difference between "gene kinds" and "gene species" are nominal, and the metagenome that a microbiota carries is essentially a 'community' of metagenomic genes (MGs). Each row of a MGA table represents a metagenome of a microbiota, and the whole MGA table represents a 'meta-metagenome' (or an assemblage of metagenomes) of *N* microbiotas (microbiome samples). Consequently, the same ecological/network analyses used in OTU analyses should be equally applicable to MGA tables. Here we choose to analyze the heterogeneity of metagenome by introducing classic Taylor's power law (TPL) and its recent extensions in community ecology. Heterogeneity is a fundamental property of metagenome, particularly in the context of human microbiomes. Recent studies have shown that the heterogeneity of human metagenomes is far more significant than that of human genomes. Therefore, without deep understanding of the human metagenome heterogeneity, personalized medicine of the human microbiome-associated diseases is hardly feasible. The TPL extensions have been successfully applied to measure the heterogeneity of human microbiome based on amplicon-sequencing reads of marker genes (e.g., 16s-rRNA). In this article, we demonstrate the analysis of the metagenomic heterogeneity of human gut microbiome at whole metagenome scale (with type-I power

OPEN ACCESS

Edited by:
Hyun-Seob Song, University of Nebraska–Lincoln, United States

Reviewed by:
Mina Rho, Hanyang University, South Korea
Jennifer Mobberley, Pacific Northwest National Laboratory (DOE), United States

***Correspondence:**
Zhanshan (Sam) Ma
ma@vandals.uidaho.edu

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *01 July 2019*
Accepted: *20 March 2020*
Published: *06 May 2020*

Citation:
Ma ZS (2020) Assessing and Interpreting the Metagenome Heterogeneity With Power Law. Front. Microbiol. 11:648. doi: 10.3389/fmicb.2020.00648

law extension) and metagenomic gene scale (type-III), as well as the heterogeneity of gene clusters, respectively. We further examine the influences of obesity, IBD and diabetes on the heterogeneity, which is of important ramifications for the diagnosis and treatment of human microbiome-associated diseases.

Keywords: metagenome ecology, metagenomic gene abundance (MGA) table, Taylor's power law, power law extensions, metagenome spatial heterogeneity, metagenome functional gene cluster (MFGC), medical ecology of metagenome

INTRODUCTION

Understanding the microbiome or "the biome of microbes" usually starts with cataloging the list of OTUs (operational taxonomic units) and tabulating their abundance distribution, leading to the so-termed OTU table. The OTU table has a counterpart in macrobial ecology of plants and animals, known as species abundance distribution (SAD). The recognition of the equivalence between OTU table (or OTU distribution) and SAD has greatly facilitated the infiltration of macrobial ecology theories into microbial ecology. The translation and testing of the ecological theories originated in macrobial ecology with microbiome datasets also lead to the ongoing development of a *unified* or *inclusive* ecology of plants, animals and microbes. Of course, OTU tables, which are usually obtained through amplicon sequencing of marker genes (e.g., 16S-rRNA for bacteria or 18S-rRNA for fungi), are not sufficient for understanding microbiome. For this reason, scientists investigate the metagenome (i.e., the total genomes of all microbes in a microbiome) by using the whole-genome or metagenome shotgun (MGS) sequencing technology. The output from the MGS sequencing technology is the metagenomic gene abundance (MGA) table, which is rather similar to the OTU table, given that both are the abundance of genes (i.e., 16S-rRNA gene *vs*. regular genes). Nevertheless, there is an essential difference between the OTU table and MGA table: the OTU table carries taxonomic information, but MGA table carries genetic or gene information. The former has been a *de fact* standard entity in ecological analyses of the microbiome datasets, and the latter has been mostly used in genetic and evolutionary analyses. In existing metagenomic research, however, few ecological analyses have been performed with MGA data. We argue that the ecological analysis of metagenomic MGA, or "the ecology of metagenome," is an emerging field where ecological theories should play a critical role.

The similarity between OTU and MGA tables is far from superficial. The familiar OTU table is a matrix of OTU *reads* that capture the species abundance distribution (SAD) of all species in N microbial communities (e.g., N microbiome samples from N individuals, spatial sites or time-points of an individual), with each row corresponding to the SAD of each species, which is simply the frequency distribution (relative abundance) of an OTU across N samples. Together, an OTU table represents a meta-community or ecosystem (when meta-factors were added as special columns) in terms of species abundance distribution, including both taxonomic identities and their population abundances in the system. Various ecological analyses (theories and models) such as diversity analysis, power law, diversity-area relationship (DAR), neutral theory and network analyses have been conducted with OTU tables, to reveal important insights on the structure, dynamics and functions of microbiomes (e.g., Costello et al., 2012; Lozupone et al., 2012; Hanson et al., 2012; Human Microbiome Project Consortium [HMP], 2012; Barberaìn et al., 2014; Ma, 2015, 2018, 2019; Ma and Li, 2018, 2019; Li and Ma, 2019; Ma and Ellison, 2019; Ma et al., 2019). These analyses have become a *de facto* standard for 16S-rRNA based (amplicon-sequencing based) microbiome research. However, few such analyses have been applied to MGA tables.

Conceptually, if we conceive the metagenomic genes as "gene species," then these gene species or genes (we use both the terms interchangeably hereafter) constitute "a community of gene species," which is essentially the concept of *metagenome*. Each metagenome constitutes one row of a MGA table. In other words, a MGA table consists of multiple (N) metagenomes, corresponding to N metagenome samples, and a MGA table can be considered as an assemblage or meta-community of metagenomes. Here we coin the term "*assemblage of metagenomes*" (=*metagenome assemblage*) or "*assemblage*" (when no confusion occurs) to represent "metacommunity of metagenomes" or 'meta-metagenome' and also to avoid the double prefix of 'meta-'. Therefore, a MGA table represents an assemblage of metagenomes, consisting of N metagenomes, e.g., from N individuals (or samples). When meta-factors (such as host physiology) are added to a MGA table, then the MGA table describes an "ecosystem of metagenomes." With such conceiving, we argue that ecological and network analyses can be harnessed to investigate important problems in metagenome research such as diversity (Ma and Li, 2018), heterogeneity, functional redundancy, mechanisms of diversity maintenance, inter-gene interactions, and dynamics of metagenomes. In a previous study, we successfully demonstrated the application of Hill numbers for measuring metagenome diversity and similarity (Ma and Li, 2018). In this study, we demonstrate the application of Taylor's power law (Taylor, 1961, 1984; Taylor and Taylor, 1977) and its recent extensions to community ecology (Ma, 2015) to assess and interpret the *heterogeneity* of metagenome assemblage.

According to Li and Reynolds (1995) *heterogeneity* can be defined based on two components: the system property of interests and its complexity or variability. They defined heterogeneity as "the complexity and/or variability of a system property in space and/or time" (Li and Reynolds, 1995). To some extent, considering heterogeneity as the other side of evenness coin or as a proxy of biodiversity is not

unreasonable. However, if we look into its usage in population ecology, specifically in the studies on the population spatial distribution of animal or plants, we may quickly recognize one significant difference in community heterogeneity and community diversity. That is, the former is either explicitly or implicitly associated with certain spatial elements, but the latter is not, arguably, beta-diversity is an exception. In addition, heterogeneity is a "group" property in the sense that comparing heterogeneity generally requires at least two entities. As a side note, the heterogeneity in time (states) or temporal heterogeneity is similar to (temporal) stability (Ma, 2015) and is not a topic of this study. In the following, we use the term *heterogeneity* to refer to *spatial heterogeneity* whenever confusion is unlikely.

In the following, we demonstrate the assessment and interpretation of the metagenomic heterogeneity of human gut microbiome at whole metagenome scale with type-I power law extension (PLE) and metagenomic gene scale (type-III PLE), as well as the heterogeneity of functional gene clusters, respectively. Here, the term *spatial* can be applied to different individuals or different microbiome habitats of an individual in the case of human microbiomes, or samples from different habitats in the case of general environmental metagenomes. Furthermore, we also investigate the influence of three common microbiome associated diseases (obesity, diabetes, and IBD) on the metagenomic spatial heterogeneity in human gut systems.

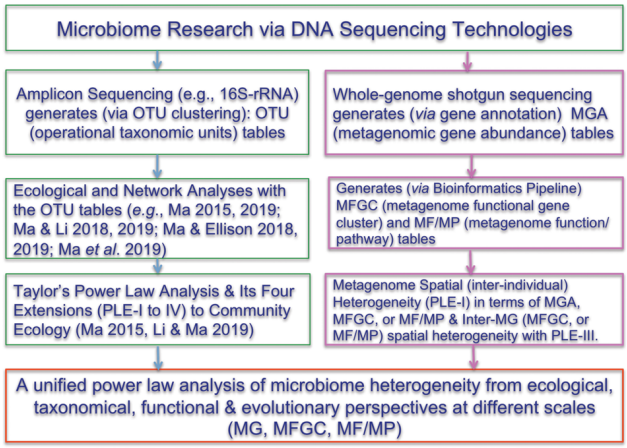


FIGURE 1 | Showing the flowchart of analyzing the microbiome heterogeneity from ecological, taxonomical, functional and evolutionary perspective in terms of various scales [OTU, MG (metagenomic gene), MFGC (metagenome functional gene clusters), MF/MP (metagenomic function/pathway) with the power law extensions (PLEs)]. The right side and framed in red color are newly introduced in the present study. See the Online **Supplementary Information** (OSI) for the R-Scripts implementing the PLE analysis and randomization tests.

CONCEPTS AND DEFINITIONS

One of the most important findings that the Human Microbiome Project Consortium [HMP] (human microbiome project) has revealed is the enormous inter-subject difference or heterogeneity among individual subjects. However, much of the evidence supporting the notion of personalized microbiome comes from 16S-rRNA datasets. This is because the OTU tables generated from 16S-rRNA sequencing are inherently more submissive to ecological analyses than the MGA tables generated from the whole-genome metagenomic sequencing are. Indeed, compared with the analysis of 16S-rRNA OTU tables, the applications of ecological theories (laws) to the metagenome MGA data analysis have been much fewer. Here, we propose to introduce Taylor's power law (Taylor, 1961, 1984, 2007; Taylor and Taylor, 1977; Taylor et al., 1983, 1988) and its recent extensions (Ma, 2015; Oh et al., 2016) to the ecological community, for assessing and interpreting the spatial (or inter-subject) heterogeneity within the metagenome assemblage represented by a MGA table. **Figure 1** below shows the flowchart of various ecological and bioinformatics analyses involved in the present study.

Taylor's (1961) power law, describing the scaling relationship between the population mean abundance (m) and its variance (V) over space (i.e., $V = am^b$), is one of few well recognized ecological laws in population ecology, and it offers a powerful mathematical tool to measure the spatial *aggregation* (*heterogeneity*). Its power law scaling parameter (b) often embodies rich ecological and evolutionary insights about specie abundance and distribution over space or time across different environments (Taylor, 1961, 1984, 2007; Taylor and Taylor, 1977; Taylor et al., 1983, 1988). Since its discovery more than a half century ago (Taylor, 1961), Taylor's power law has been the target of numerous field tests and theoretical analyses, especially in macrobial ecology of plants and animals. In particular, a resurgence of theoretical investigation and extensions to even wider applications in many fields of science and technology, particularly inter-disciplinary studies, has been ongoing in the last few years (e.g., Reuman et al., 2009, 2014, 2017; Cohen et al., 2012, 2013; Ma, 2012, 2015; Stumpf and Porter, 2012; Wearn et al., 2013; Cohen, 2014; Zhang et al., 2014; Cohen and Xu, 2015; Cohen and Saitoh, 2016; Oh et al., 2016; Tippett and Cohen, 2016; Quist et al., 2017). In a previous study (Ma, 2015), we extended the original Taylor's (1961) power law from population to community level and tested four power law extensions (PLEs) with the 16s-rRNA amplicon-sequencing datasets of the microbial communities from the human microbiome project (Human Microbiome Project Consortium [HMP]). Among the four PLEs introduced by Ma (2015), Type-I and Type-III PLEs can quantify the community (level) spatial heterogeneity and mixed-species (level) spatial heterogeneity, respectively. Type-II and Type-IV were proposed to assess the community temporal stability and mixed-species temporal stability, respectively, but this study does not implicate them since both Type-II and IV require time-series data, for which we did not get sufficiently large datasets, but they should still be applicable for measuring the metagenome stability.

PLE-I (Type-I Power Law Extension) for Measuring Metagenome Spatial Heterogeneity

Similar to the PLE-I for measuring community spatial heterogeneity (Ma, 2015), we propose to use the following mean-variance power function for measuring the *metagenome spatial heterogeneity* of a metagenome assemblage (or meta-metagenome, as explained previously):

$$V_s = am_s^b \quad s = 1, 2, \ldots, S \tag{1}$$

where m_s is the mean of gene abundances of all genes (species) (G) in the metagenome of an individual subject (s = 1, 2, ...S), V_s is the corresponding variance, S is the number of subjects, and G is the number of genes contained in the metagenomes of the S subjects. Note m_s is the mean gene abundance *per* gene species, not per subject, which is different from the case in PLE-III (type-III power law extension) introduced below. In addition, the fitting of Eqn. (1) is performed with S data points, i.e., across S individual subjects (or S metagenomes), rather than across genes, as in the case of PLE-III below.

The parameter b describes the *fractional scaling* of V-m relationship statistically, or the *metagenome spatial heterogeneity* biologically. When b = 1, the heterogeneity is random, which means that the heterogeneity—the inter-subject difference in their gene abundance distribution—is essentially random, statistically follows Poisson distribution. When $b > 1$, the inter-subject heterogeneity is non-random and follows highly skewed long-tail distribution (such as the power law distribution). When $b < 1$, the inter-subject heterogeneity in their metagenome is fixed, or follows the uniform statistical distribution. From field studies in ecology, the cases when $b = 1$ or $b < 1$ are extremely rare in real world and usually only exist theoretically (Taylor, 1961, 1984). We term a metagenome assemblage (i.e., an assemblage of metagenomes) with $b = 1$ random metagenome (strictly speaking, metagenome assemblage), $b < 1$, homogenous metagenome, and $b > 1$, heterogeneous metagenome.

Parameter a in Eqn. (1) is meanwhile related to sampling related factors such as sampling unit or sequencing platforms, but is little influenced by biological interactions. Hence, we generally do not attempt to draw biological interpretations from parameter a due to the strong influence from sampling. It is noted that parameter a also has the same interpretation in PLE-III below.

We further define *critical diversity of metagenome heterogeneity* (CDMH) or m_0 as:

$$m_0 = \exp[\ln(a)/(1-b)] \quad (b \neq 1) \tag{2}$$

where a and b are PLE-I parameters from eqn. (1). The CDMH or m_0 is the *mean gene abundance* level (per gene species) at which metagenome spatial heterogeneity is random, and across which the heterogeneity transits to either heterogeneous (when $m > m_0$) or regular (uniform or fixed) (when m $< m_0$). Since the *mean gene abundance*, although termed abundance, is essentially a measure of *gene diversity* (i.e., the mean abundance of various gene species in a metagenome), we used the term *critical diversity of metagenome heterogeneity*, rather than using the term "*critical abundance*." The latter is indeed used in the next section for PLE-III, which is the average of single gene abundances from various individuals and consequently the term *abundance* is more appropriate.

PLE-III (Type-III Power Law Extension) for Measuring Gene-Level Spatial Heterogeneity

Similar to the PLE-III for measuring mixed-species spatial heterogeneity in Ma (2015), we propose to use the following mean-variance power function for measuring the *gene-level* (*inter-gene,* or *mixed-gene*) *spatial heterogeneity*:

$$V_g = am_g^b \quad g = 1, 2, \ldots, G \tag{3}$$

where m_g is the mean abundance of *g-th* gene, averaged across S subjects (g = 1, 2, ...G), V_g is the corresponding variance, and G is the number of gene kinds (gene species), and S is the number of individual subjects sampled. Note m_g is the mean gene abundance of the g-th gene species *per* subject (not *per* gene), which is opposite from the case in the previously introduced PLE-I for measuring metagenome spatial heterogeneity. In addition, the fitting of Eqn. (3) is performed with G data points, i.e., across all G gene species, rather than across S subjects, as in the case of the previous PLE-I.

Note that the notion of "mixed-gene" is similar to the concept of mixed-species population in the original Taylor's power law (Taylor and Woiwod, 1982; Taylor, 1984). It refers to a virtually "averaged assemblage" of genes, in which the identities or kinds of different genes were ignored. The m-V pairs are regressed (see below, through log-linear transformation into linear regression) across multiple gene species (millions in the case of this study) in a *mixture* manner. Given that the notion of gene *species* is not widely used in metagenomic research, we suggest using the term *gene-level* or *inter-gene* heterogeneity, rather than *mixed-gene* heterogeneity in the context of PLE-III.

When b = 1, the heterogeneity among metagenomic genes in terms of their gene abundance distributions should be random, i.e., all genes in the metagenome are equivalent to each other in terms of their abundance distribution, similar to the neutrality assumption in the neutral theory of biodiversity. When $b < 1$, the heterogeneity or difference among genes, if any, should be fixed, or follow a uniform distribution statistically. Both the cases of $b = 1$ or $b < 1$ should be extremely rare in real world, and are mostly theoretical possibilities. In practice, $b > 1$ should be the norm rather than the exception for metagenome heterogeneity at the gene-level. When $b > 1$, we say the metagenome is heterogeneous or aggregated in terms of its *gene-level* or *inter-gene* heterogeneity.

We further define critical abundance of gene-level heterogeneity (CAGH) or m_0 as:

$$m_0 = \exp[\ln(a)/(1-b)] \quad (b \neq 1) \tag{4}$$

where a and b are PLE-III parameters from eqn. (3). The CAGH or m_0 is the mean gene abundance level (per individual or

sample) at which gene-level spatial heterogeneity is random, and across which the heterogeneity transits to either heterogeneous (when $m > m_0$) or regular (uniform or fixed) (when $m < m_0$).

Statistical Fitting of PLE-I or PLE-III

To fit the power law, including PLE-I and PLE-III, the most commonly used approach is to transform the power law model [eqn. (1) or (3)] into the following linear function:

$$\ln(V) = \ln(a) + b\ln(m) \quad (5)$$

where all the variables (m, V) and parameters (a, b) have the exactly same interpretations as those in eqn. (1) or eqn. (3). Standard linear regression procedure can be applied to fit the model. In fact, there is an advantage for adopting the simple linear transformed regression approach, which is related to an important property of power law, scale-invariance. This property makes parameter a less relevant for determining the most important parameter of power law, i.e., the scaling parameter b (Ma, 2015). It is for this reason that we choose the simple linear regression approach for fitting all the power law models. This allows us to focus on the scaling parameter (b) for assessing and interpreting the metagenome heterogeneity revealed by the metagenomic sequencing data.

As a side note, we may define *metagenome temporal stability* with Type-II PLE or *gene-level temporal stability* with Type-IV PLE, similar to Ma (2015) for community temporal stability or mixed-species temporal stability, but their demonstrations require time-series MGA data obtained with metagenomic (whole-genome or shotgun) sequencing technologies. We failed to find sufficiently long time-series MGA data to demonstrate the PLE-II or PLE-IV models and won't further discuss the temporal versions of the PLE in this study (Nielsen et al., 2014).

Bioinformatics Analysis of Metagenomic Sequencing Data

To fit the power law model, one has to first compute MGA tables from metagenomic sequencing raw reads (also known as shotgun or whole-genome sequencing) by using standard bioinformatics software pipelines (e.g., Li and Godzik, 2006; Qin et al., 2010, 2012; Zhu et al., 2010; Chatelier et al., 2013; Li et al., 2014; Xiao et al., 2015, 2016; Wang and Jia, 2016; Sczyrba et al., 2017; Ma and Li, 2018).

Millions of contigs are obtained through the metagenome assembly step. Those millions of contigs are fed into *gene prediction* software and the latter generate a list of non-redundant genes based on the criteria set by ORFs (open reading frames). We term those non-redundant genes as *metagenomic genes* (MGs) or simple genes. MG embodies single-gene-level genetic information, and its number in a typical metagenome sample is in the magnitude of millions (Ma and Li, 2018). The previously defined MGA table is actually the table of MGs.

Directly characterizing or summarizing information from the millions of MGs or MGA tables can be rather challenging. An alternative research strategy is to first group those millions of genes (MGs) into functional gene clusters, and then investigate the properties of the *functional gene clusters*. There are mature bioinformatics algorithms and software pipelines to cluster the millions of MGs into hundreds of MFGCs (metagenome functional gene clusters), and the magnitude of MFGC numbers (hundreds) is much small that that of the MGs (millions) (Ma and Li, 2018). Obviously, the huge reduction in the magnitudes from MGs (millions) to MFGCs (hundreds) should make our measuring metagenomic spatial heterogeneity simpler.

DEMONSTRATION AND DISCUSSION

The Datasets of Metagenomes

We collected three gut metagenome datasets from public domain including, 264 stool samples from overweight and lean individuals (Qin et al., 2010; Chatelier et al., 2013), 145 stool samples from type-2 diabetes and healthy controls (Qin et al., 2012), and 219 stool samples from IBD patients and healthy controls. A total of 628 metagenome samples with their metagenomic gene (MG) catalog and the gene abundance (MGA) tables for each dataset were computed with standard metagenomic analysis pipelines (e.g., Li and Godzik, 2006; Qin et al., 2010, 2012; Zhu et al., 2010; Chatelier et al., 2013; Li et al., 2014; Xiao et al., 2015, 2016; Wang and Jia, 2016; Sczyrba et al., 2017). Furthermore, we defined metagenome functional gene clusters (MFGC) based on Ma and Li (2018) and obtained their abundance tables. **Supplementary Table S1** showed more detailed information about the three datasets we use in this paper for demonstrating the application of the power law.

Specifically, after whole-genome (shotgun) sequencing of a metagenome sample, sequencing reads from the fecal samples were processed for quality control, removal of human sequences, assembling, assembly revision and gene prediction by using MOCAT pipeline (Kultima et al., 2012). This pipeline consists of a series of software packages, which can process metagenomes in a standardized and automated manner while improving the quality of assembly and gene prediction at run time. In the pipeline, FASTX Toolkit[1] was used for quality control; SOAPaligner2 (Li et al., 2009) for identifying human sequences; SOAPdenovo v1.06 (Li et al., 2010) for assembling; MetaGeneMark (Zhu et al., 2010) for gene prediction; CD-HIT (Li and Godzik, 2006) for clustering genes in each cohort.

The details of the data/software/parameters used to compute the MGA tables can be found in the online method of Li et al. (2014). In fact, the MGA tables are available online at: http://meta.genomics.cn/meta/dataTools. Li et al. (2014) annotated the metagenomic genes according to the "Kyoto Encyclopedia of Genes and Genomes" (KEGG) and the "evolutionary genealogy of genes non-supervised orthologous groups" (eggNOG) databases. They further identified a total of 6,980 KEGG orthologous groups (KOs) and 36,489 eggNOG orthologous groups, accounting for 51.6 and 69.3% of the total sequencing reads.

[1] http://hannonlab.cshl.edu/fastx_toolkit/

TABLE 1 | The parameters of PLE-I (type-I power law extension) for *metagenome spatial heterogeneity*, in terms of the MGA (metagenomic gene abundance).

Power Law Extension (PLE)	Case study	Treatment	*b*	SE(*b*)	ln(*a*)	SE[ln(*a*)]	m_0	*R*	*p*-value	*N*
Type-I PLE for Metagenome Spatial Heterogeneity with MGA	Obesity	Lean	2.012	0.113	3.740	0.337	0.025	0.878	<0.001	95
		Overweight	3.447	0.158	−1.204	0.532	1.636	0.914	<0.001	96
	Type-II Diabetes	Healthy	3.232	0.210	−0.529	0.650	1.267	0.876	<0.001	74
		Disease	1.846	0.143	3.982	0.447	0.009	0.840	<0.001	71
	IBD	Healthy	1.385	0.079	5.365	0.266	0.000	0.903	<0.001	71
		Disease	2.248	0.227	2.754	0.761	0.110	0.766	<0.001	71

In the online **Supplementary Information** (OSI), the R-Scripts for implementing the power law analysis and randomization tests for determining the differences in the PLE parameters are provided.

Metagenome Spatial Heterogeneity in Terms of the MGA (Metagenomic Gene Abundance) Spatial (Inter-Subject) Distribution Measured With PLE-I

We first fitted PLE-I (type-I power law extension) with metagenomic gene abundance (MGA) datasets directly in order to measure the metagenome spatial (inter-subject) heterogeneity for each treatment (group) of the three datasets, and the results were listed in **Table 1**, from which we summarize the following findings:

(i) PLE-I fitted to all three datasets extremely well with p-value < 0.0001. This indicates the ubiquitous applicability of the PLE for assessing the metagenome spatial (i.e., inter-subject) heterogeneity of either MGA (this section) or MFGC (the next section).

(ii) The scaling parameter (b) of PLE-I for the most treatments is between *2* and *4* except for the two treatments (diseased treatment in the diabetes study, and the healthy treatment in the IBD study), and the parameter (b) varied significantly between the treatments with a range of [*1.385, 3.447*].

(iii) The values of the scaling parameter (b) for the healthy samples (group) and diseased samples (group) were significantly different (p-value < 0.05), in all three case studies (obesity, diabetes and IBD). Therefore, we conclude that PLE-I can be harnessed to measure the *metagenome spatial heterogeneity* in terms of gene abundance distribution. Furthermore, it has a potential being a discriminant metric for distinguishing between the healthy and diseased metagenome samples, as revealed in **Table 2** (p-value < 0.05), in which randomization test (Collingridge, 2013) with 1000 times of re-sampling was utilized to test the difference in the b-value between the healthy and diseased treatments. **Figure 2** shows the fitted power law models for the obesity case study, i.e., one straight line for the lean group and another for the overweight group.

TABLE 2 | The p-value of the randomization test for the difference between the healthy and diseased treatments in their metagenome spatial heterogeneities parameters of PLE-I.

Power Law Extension (PLE)	Case Study	Treatments	*b*	ln(*a*)	m_0
Type-I PLE for Metagenome Spatial Heterogeneity with MGA	Obesity	Lean vs. Overweight	<0.001	<0.001	<0.001
	Type-2 diabetes	Healthy vs. Disease	0.044	0.038	0.044
	IBD	Healthy vs. Disease	0.021	0.043	0.015

The *metagenome spatial heterogeneity* is the counterpart of *community spatial heterogeneity* in community ecology, and it measures the spatial heterogeneity of metagenomes of individual subjects or inter-subject metagenome heterogeneity in a population (or cohort), similar to measuring the heterogeneity among spatially explicit local communities in community ecology (Ma, 2015). With our newly coined term of metagenome assemblage, parameter b measures the heterogeneity of metagenome assemblage represented by a MGA table. The higher b-value of PLE-I represents greater heterogeneity (unevenness or diversity) among individuals in their metagenomes in terms of their gene abundance distributions. When $b = 1$, it implies that the differences among individuals are random. When $b < 1$, it implies that the differences among individuals follow *uniform* distribution statistically (i.e., a fixed difference).

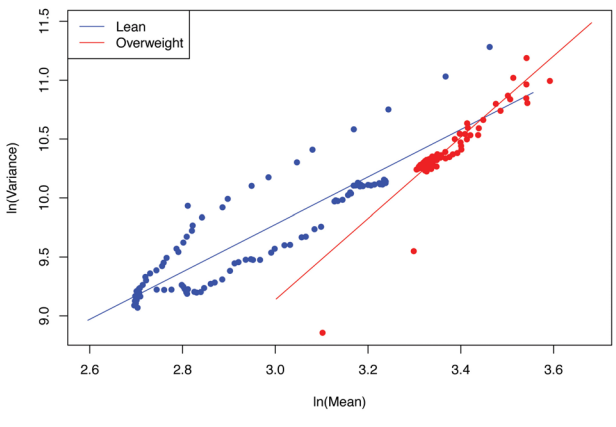


FIGURE 2 | The PLE-I (type-I power law extension) models fitted for the obesity case study.

Gene-Level (or Inter-Gene) Spatial Heterogeneity in Terms of the MGA (Metagenomic Gene Abundance) Distribution Measured With PLE-III

We also fitted the PLE-III (type-III power law extension) with metagenomic gene abundance (MGA) datasets directly in order to measure the gene-level or mixed-gene spatial heterogeneity for each of the 3 datasets, and the results were listed in **Table 3** below. It was shows that:

(i) The PLE-III fitted to all 3 datasets extremely significant with p-value < 0.0001, and the standard errors of the model parameters were close to zero. The linear correlation coefficients were between 0.949 and 0.961. All the criteria indicate that the goodness-of-fitting to PLE-III was extremely well given millions of data points were fitted.

(ii) The parameter b of PLE-III for all the treatments fall in a rather narrow range of [2.340, 2.466]. Therefore, we conclude that PLE-III can be harnessed to measure the gene-level spatial heterogeneity in terms of the gene abundance distribution, but its application for discriminating the healthy and diseased samples is of limited value given its insensitivity to host factors such as diseases. **Figure 3** shows the fitted PLE-III with the dataset from the obesity study.

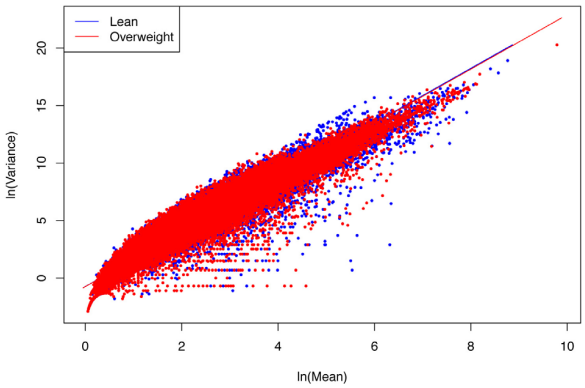


FIGURE 3 | The PLE-III (type-III power law extension) models fitted to the obesity study datasets: more than 10 million points (5407291 lean group + 5134721 overweight) were used to fit the PLE-III models, but here we only randomly selected 100,000 points (50,000 from each treatment) to draw the graphs (so as to accommodate the file size of the figure).

Taylor's power law has been tested with hundreds, if not thousands, of field studies and many theoretical examinations (Taylor, 1961, 1981, 1984, 2007; Taylor et al., 1983, 1988; Ma, 1991, 2012, 2013, 2015, 2018; Reuman et al., 2009, 2014, 2017; Stumpf and Porter, 2012; Cohen and Xu, 2015; Tippett and Cohen, 2016; Quist et al., 2017). However, to the best of our knowledge, the tests exhibited in **Table 3** should be the cases that have used the biggest numbers of data points (the column N in **Table 3**) to fit the power law model, since it was first discovered more than a half century ago. For example, in the case of obesity study, for each of the two treatments (lean *vs*. overweight), more than five million genes were used to fit PLE-III model. This shows the exceptional robustness of the power law model.

The PLE-III for measuring the gene-level or mixed-gene spatial heterogeneity is the counterpart of mixed-species spatial aggregation in community ecology (Ma, 2015). The term *aggregation* is often used in population ecology, and it is the counterpart of heterogeneity in community ecology. As explained previously, the term *mixed-gene* setting assumes that we ignore the identities of individual genes, and what is measured is the aggregation (unevenness or heterogeneity) of an *average gene* species. We suggest using the term "*gene-level spatial heterogeneity*" for what is measured with the PLE-III in metagenomic research.

Metagenome Spatial Heterogeneity in Terms of the MFGC (Metagenome Functional Gene Cluster) Distribution Measured With PLE-I

According to Ma and Li (2018), the term MFGC (metegenome functional gene cluster) refers to cluster of functionally similar or same genes, generated from functional annotation or gene annotation through online mapping to functional databases such as KEGG (for metabolic pathways) and eggNOG (for protein functions). Hence, MFGC is purely functionality-based and is mostly cross-species. One of its unique advantages is that it embodies the functional redundancy in microbiome very well. The difference between Type-I MFGC (MFGC-I) and Type-II MFGC (MFGC-II) lies in their differences in handling the genes within each cluster. In MFGC-I, only the number of gene species (kinds) is counted but the abundance of individual gene is ignored. In MGGC-II, both the number of gene species (kinds) and the abundance of each gene matter in the analysis. In other

TABLE 3 | The parameters of PLE-III (type-III power law extension) for measuring gene-level (inter-gene) spatial aggregation, in terms of the metagenomic gene abundance (MGA).

Power Law Extension (PLE)	Case study	Treatment	b	SE(b)	ln(a)	SE[ln(a)]	m_0	R	p-value	N
Type-III PLE for Gene-Level Spatial Heterogeneity with MGA	Obesity	Lean	2.371	0.000	−0.732	0.001	1.706	0.961	<0.001	5407291
		Overweight	2.363	0.000	−0.744	0.001	1.726	0.961	<0.001	5134721
	Type-II Diabetes	Healthy	2.340	0.000	−0.842	0.001	1.875	0.954	<0.001	4573927
		Disease	2.338	0.000	−0.791	0.001	1.806	0.949	<0.001	4432814
	IBD	Healthy	2.466	0.000	−1.000	0.001	1.978	0.961	<0.001	2898618
		Disease	2.351	0.000	−0.791	0.001	1.796	0.957	<0.001	4462890

words, with MFGC-I, we only care the *number* of gene *species* (kinds), and with MFGC-II we care both the number of *gene species* (*kinds*) and the abundance of each gene within each cluster. This treatment is very similar to a common practice in community ecology, where a simplified measure for biodiversity is to only count the number of species (also known as *species richness*), and a more comprehensive measure of biodiversity uses more sophisticated entropy such as Shannon entropy, which consider both species richness and abundances.

In the previous section, we conducted power law analysis in terms of the metagenomic gene (MG) or metagenomic gene abundance (MGA) distribution. In this section, our analysis is performed in terms of the metagenome functional gene cluster (MFGC). That is, using the *MFGC abundance tables* (similar to MGA or OTU tables, except that the entity is the MFGC) to fit the PLE models.

The results of fitting the PLE-I with MFGC tables were listed in **Table 4**, from which we can observe the following findings:

(i) The PLE-I model fitted to the MFGC abundances extremely well (significant with p-value < 0.0001), and this indicates the ubiquitous applicability of the PLE for assessing the metagenome spatial (i.e., inter-subject) heterogeneity of either MFGCs (this section) or MG (previous section).

(ii) MFGC-I and MFGC-II exhibited slightly different scaling parameter (b) values. The scaling parameter b of PLE-I ranged [2.027, 2.138] for MFGC-I and [1.715, 1.988] for MFGC-II, indicating that the MFGC-I has a higher heterogeneity degree. This difference should be due to their definitional difference: MFGC-I ignored the information of individual gene abundances, only taking into account the number of gene species (kinds), or gene *richness*. Obviously, ignoring the gene abundance information should lead to larger heterogeneity (difference), which explains why the PLE-I parameter b of MFGC-I was slightly higher ($b > 2$), while that of MFGC-II was lower ($b < 2$). Furthermore, MFGC-II should better embody functional redundancy information given that it considers both gene species (kinds or richness) and abundances.

(iii) Although MFGC-I and MFGC-II displayed slightly different ranges in their parameter b, the b-values from two databases (eggNOG and KEGG) within each MFGC type were rather close with each other and the difference was negligible. This simply indicates that the heterogeneity scaling based on metabolic pathways (KEGG database) or protein functions (eggNOG) makes little difference. This should be expected given that, at the MFGC level, both eggNOG and KEGG should be controlled by the same underlying gene-level mechanisms.

(iv) As shown in **Table 5**, in terms of the parameter changes associated with diseases, only IBD treatment displayed significant difference from its healthy control in MFGC-I, and the other diseases treatments did not exhibit any significant differences from their healthy controls. This result suggests that at the MFGC level, the metagenome spatial heterogeneity is less sensitive to diseases than at the

TABLE 4 | The parameters of PLE-I (type-I power law extension) for metagenome spatial heterogeneity, in terms of the MFGC (metagenome functional gene cluster) distribution.

Type of MFGC and database used	Microbiome	Treatment	b	$SE(b)$	$\ln(a)$	$SE[\ln(a)]$	R	p-value	N	m_0
Type-I MFGC (eggNOG)	Obesity	Lean	2.119	0.020	3.187	0.157	0.996	<0.0001	95	0.058
		Overweight	2.028	0.017	3.895	0.135	0.997	<0.0001	96	0.023
	Type 2 diabetes	Healthy	2.058	0.025	3.501	0.180	0.995	<0.0001	74	0.037
		Disease	2.057	0.018	3.480	0.126	0.997	<0.0001	71	0.037
	IBD	Healthy	2.053	0.017	3.690	0.136	0.998	<0.0001	71	0.030
		Disease	2.138	0.021	3.014	0.159	0.997	<0.0001	71	0.071
MFGC Type-I (KEGG)	Obesity	Lean	2.091	0.015	3.505	0.123	0.998	<0.0001	95	0.040
		Overweight	2.027	0.014	4.008	0.109	0.998	<0.0001	96	0.020
	Type-II diabetes	Healthy	2.035	0.020	3.772	0.150	0.997	<0.0001	74	0.026
		Disease	2.036	0.014	3.794	0.106	0.998	<0.0001	71	0.026
	IBD	Healthy	2.042	0.013	3.888	0.104	0.999	<0.0001	71	0.024
		Disease	2.111	0.016	3.292	0.130	0.998	<0.0001	71	0.052
MFGC Type-II (eggNOG)	Obesity	Lean	1.884	0.021	4.912	0.223	0.995	<0.0001	95	0.004
		Overweight	1.859	0.019	5.212	0.208	0.995	<0.0001	96	0.002
	Type-II diabetes	Healthy	1.783	0.059	5.771	0.614	0.962	<0.0001	74	0.001
		Disease	1.715	0.091	6.461	0.937	0.915	<0.0001	71	0.000
	IBD	Healthy	1.967	0.024	3.992	0.260	0.995	<0.0001	71	0.016
		Disease	1.937	0.021	4.295	0.232	0.996	<0.0001	71	0.010
MFGC Type-II (KEGG)	Obesity	Lean	1.915	0.017	4.834	0.197	0.996	<0.0001	95	0.005
		Overweight	1.889	0.016	5.136	0.178	0.997	<0.0001	96	0.003
	Type-II diabetes	Healthy	1.830	0.054	5.572	0.577	0.970	<0.0001	74	0.001
		Disease	1.806	0.078	5.856	0.831	0.942	<0.0001	71	0.001
	IBD	Healthy	1.988	0.021	3.997	0.234	0.996	<0.0001	71	0.017
		Disease	1.961	0.018	4.257	0.201	0.997	<0.0001	71	0.012

TABLE 5 | The *p*-value of the randomization test for the difference between the healthy and diseased treatments in their PLE-I (type-I power law extension) parameters in terms of the MFGC.

MFGC Type and Databases used	Microbiome	Treatments	*b*	ln(*a*)	m_0
MFGC Type-I (eggNOG)	Obesity	Lean vs. Overweight	0.347	0.345	0.348
	Type 2 diabetes	Healthy vs. Disease	0.985	0.937	0.965
	IBD	Healthy vs. Disease	0.039	0.033	0.059
MFGC Type-I (KEGG)	Obesity	Lean vs. Overweight	0.442	0.465	0.444
	Type 2 diabetes	Healthy vs. Disease	0.987	0.913	0.947
	IBD	Healthy vs. Disease	0.018	0.012	0.025
MFGC Type-II (eggNOG)	Obesity	Lean vs. Overweight	0.421	0.388	0.432
	Type 2 diabetes	Healthy vs. Disease	0.551	0.556	0.597
	IBD	Healthy vs. Disease	0.330	0.370	0.375
MFGC Type-II (KEGG)	Obesity	Lean vs. Overweight	0.361	0.337	0.382
	Type 2 diabetes	Healthy vs. Disease	0.781	0.771	0.787
	IBD	Healthy vs. Disease	0.370	0.427	0.418

TABLE 6 | The *p*-value of Wilcoxon tests for the difference between the healthy and diseased treatments in their metagenome spatial heterogeneities and community dominance (also see **Supplementary Figure S1A** for the V/M heterogeneity index and **Supplementary Figure S1B** for the community dominance index).

Taylor's Power Law Extension (TPLE)	Case Study	Treatments	Mean of Healthy	Mean of Diseased	*P-value*
Variance/mean-ratio heterogeneity Index (*V*/*m*)	Obesity	Lean vs. Overweight	886.50	1129.0	<0.001
	Type-2 Diabetes	Healthy vs. Disease	594.10	761.10	<0.001
	IBD	Healthy vs. Disease	786.70	1064.4	<0.001
Community dominance Index (*M**/*m*)	Obesity	Lean vs. Overweight	45.717	39.910	<0.001
	Type-2 Diabetes	Healthy vs. Disease	27.766	34.444	<0.001
	IBD	Healthy vs. Disease	28.419	37.765	<0.001

MG level, as indicated by the randomization test results in **Table 2**.

MFGC-Level (Inter-MFGC) Spatial Heterogeneity in Terms of the MFGC (Metagenome Functional Gene Cluster) Distribution Measured With PLE-III

In the previous section, we investigated the spatial heterogeneity of MFGC by using the PLE-I (type-I power law extension) modeling. That is to analyze the inter-subject heterogeneity of their metagenomes in terms of the functional gene cluster (i.e., MFGC). In this section, we investigate the spatial heterogeneity at the MFGC-level by using PLE-III (type-III power law extension). In other words, by assuming that there exists an average MFGC in a mixed-MFGC setting (by ignoring the difference among MFGCs), we assess the heterogeneity of MFGCs at the average MFGC level. Therefore, a fundamental difference between the analysis here and the analysis in the previous section is that here, the heterogeneity is measured in terms of a virtually averaged MFGC (or at MFGC-level), while in the previous section, the heterogeneity was measured in terms of the whole metagenome (or at metagenome level).

To save page space, the results for MFGC-level spatial heterogeneity were listed in **Supplementary Table S2** in the OSI (online **Supplementary Information**), from which we summarize the following findings:

(i) The PLE-III model fitted to the MFGC tables extremely significant with p-value < 0.0001 in all three case studies, and this indicates the ubiquitous applicability of the PLE for assessing the spatial (i.e., inter-subject) heterogeneity of an 'averaged' MFGC.

(ii) The scaling parameter (b) of the PLE-III model ranged narrowly [1.472, 1.654], and varied little either between the MFGC-I and MFGC-II or between the healthy and diseased treatments within each case study. This suggests that, the sensitivity of the scaling parameter (b) of PLE-III to host factors such as diseases is rather muted, and consequently may be of limited practical applications.

(iii) Contrary with the pattern of PLE-I in the previous section, where MFGC-I has slightly larger scaling parameter (b) value than MFGC-II has, here MFGC-I [1,472, 1.547] has slightly smaller b-value than MFGC-II [1.525, 1.654] does.

(iv) The scaling parameter (b) of PLE-III, estimated with KEGG or eggNOG showed little differences, similar with the previous PLE-I model.

(v) We also performed the randomization tests for the PLE-III parameters (**Supplementary Table S3**). In most cases, the model parameters did not showed significant differences between the healthy and diseased treatments.

CONCLUSION AND DISCUSSION

In previous sections, we demonstrated that PLE-I and PLE-III, originally designed to measure community spatial heterogeneity and mixed-species population spatial aggregation in community ecology (Taylor, 1961, 1984; Ma, 2015), can be introduced

to (*i*) measure *metagenome spatial heterogeneity* in terms of either MG or MFGC abundance distribution and measured with PLE-I; or (*ii*) MG-level (or MFGC-level) spatial heterogeneity measured with PL-III. The first application is a measure at the whole metagenome (more accurately, metagenome assemblage) level, because it measures the inter-subject (spatial) heterogeneity within a cohort or population of individuals in their metagenomes. The second application is a measure at MG or MFGC level, because it measures the inter-MG or inter-MFGC heterogeneity from the perspective of a virtually averaged MG or MFGC. Although we used the metagenomic datasets of the human microbiome to demonstrate the concepts and modeling analyses, the approaches should be equally applicable to the metagenomes of other microbiomes on the planet.

Traditionally, studies on heterogeneity have been mostly focused on population level, and metrics for community heterogeneity are relatively fewer. This may be to do with that community level studies are mostly focused on community diversity, instead. Nevertheless, heterogeneity and diversity are not the same (Shavit et al., 2016). First, the heterogeneity is a "group" property, while diversity can be measured with one individual or single community. Without comparing two entities, heterogeneity does not make sense. Second, diversity is a measure of numbers (and relative abundances), while heterogeneity needs to be measured by interactions and working together (Shavit et al., 2016). For example, one may say, "zoos are diverse, and natural ecosystems are heterogeneous" (Ayelet Shavit and Aaron Ellison, personal communication, 22 April 2020).

Two additional heterogeneity indexes that can be used to measure community level heterogeneity are: (*i*) Variance/mean ratio (V/m) and (*ii*) community dominance index M^*/m. Both are counterparts of spatial aggregation index and patchiness index at the population level in population ecology (Taylor, 1984; Ma, 1991). Both indexes have different definitions and interpretations. The V/m heterogeneity index is simply a ratio of the mean species abundances (m) and corresponding variance (V), and a larger index value indicates higher heterogeneity (Taylor, 1984; Ma, 1991). The community dominance index (D_c) was defined by Ma and Ellison (2018), and a larger index value indicates lower heterogeneity. **Table 6** shows the values of the computed heterogeneity indexes for the three datasets (see **Supplementary Table S1**) we used, as well as the p-values from Wilcoxon tests for the differences between the healthy and diseased treatments in each of the three datasets.

Obviously, the two heterogeneity indexes described above are much simpler to implement than the power law modeling introduced in this study. Furthermore, both indexes displayed significant differences between the healthy and diseased treatments in their metagenome heterogeneity. Given their simplicity, a natural question is: what are the advantages from using the power law modeling? The answer is that the power law analysis we presented offers tools to synthesize and measure the heterogeneity at various scales (MG, MFGC) across individuals in a cohort (population), with a unified power law model, which achieved the rare status of classic ecological laws. In fact, the power law analysis demonstrated here can also be applied to measure metagenome temporal stability at similar scales to the previous spatial versions. Furthermore, the power law analysis provides a unified modeling tool to assess and interpret the heterogeneity from ecological, taxonomical, functional and evolutionary perspectives, because it can be applied to both OTU tables and MGA tables, with the exactly same mathematical model (the power law model). When using MGA tables, it can be universally applied to the scales of the metagenomic gene or metagenome functional gene cluster.

Perhaps an even more compelling case for using the TPL/PLE parameters rather than the simple heterogeneity indexes has to do with the difference between heterogeneity and diversity. As argued previously, heterogeneity is a "group" property; measuring heterogeneity requires at least two entities. While TPL/PLE can synthesize the information from potentially unlimited number of entities (samples), the two heterogeneity indexes previously introduced were computed from single sample. To synthesize information from multiple samples, additional statistics such as the *mean* of the heterogeneity values, as displayed in **Table 6**, must be used. Nevertheless, the distribution of heterogeneity values may satisfy the Gaussian distribution. This may make the usage of mean problematic since the distribution of heterogeneity *per se* is usually highly skewed and follows power law distribution. In addition, it may even argue that the two simple heterogeneity indexes are similar to diversity measures. For example, the community dominance (M^*/m), as heterogeneity index, may even be treated as the other side of the evenness (diversity) coin (Ma and Ellison, 2018).

DATA AVAILABILITY STATEMENT

The datasets analyzed in this manuscript are publically available for downloading and the sources were noted in **Supplementary Table S1**.

AUTHOR CONTRIBUTIONS

ZM designed the study, conducted the study and wrote the manuscript.

ACKNOWLEDGMENTS

This study received funding from the following sources: A National Natural Science Foundation of China (NSFC) Grant (No. 31970116) on the Medical Ecology of Human Microbiome; The Cloud-Ridge Industry Technology Leader Award; An International Cooperation Grant (YNST) on Genomics & Metagenomics Big Data. The funders played no roles in interpreting the results.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.00648/full#supplementary-material

REFERENCES

Barberaìn, A., Casamayor, E. O., and Fierer, N. (2014). The microbial contribution to macroecology. *Front. Microbiol.* 5:203. doi: 10.3389/fmicb.2014.00203

Chatelier, E. L., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature* 500, 541–546.

Cohen, J. E. (2014). Taylor's law and abrupt biotic change in a smoothly changing environment. *Theor. Ecol.* 7, 77–86. doi: 10.1007/s12080-013-0199-z

Cohen, J. E., and Saitoh, T. (2016). Population dynamics, synchrony, and environmental quality of Hokkaido voles lead to temporal and spatial Taylor's laws. *Ecology* 97:3402. doi: 10.1002/ecy.1575

Cohen, J. E., and Xu, M. (2015). Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. *Proc. Natl. Acad. Sci. U.S.A.* 112:7749. doi: 10.1073/pnas.1503824112

Cohen, J. E., Xu, M., and Schuster, W. S. (2013). Stochastic multiplicative population growth predicts and interprets Taylor's power law of fluctuation scaling. *Proc. Biol. Sci.* 280:20122955. doi: 10.1098/rspb.2012.2955

Cohen, J. E., Xu, M., and Schuster, W. S. F. (2012). Allometric scaling of population variance with mean body size is predicted from Taylor's law and density-mass allometry. *Proc. Natl. Acad. Sci. U.S.A.* 109:15829. doi: 10.1073/pnas.1212883109

Collingridge, D. S. (2013). A primer on quantitized data analysis and permutation testing. *J. Mixed Methods Res.* 7, 79–95.

Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M., and Relman, D. A. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science* 336, 1255–1262. doi: 10.1126/science.1224203

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., and Martiny, J. B. (2012). Beyond biogeographic patterns: process shaping the microbial landscape. *Nat. Rev. Microbiol.* 10, 497–506. doi: 10.1038/nrmicro2795

Human Microbiome Project Consortium [HMP], (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Kultima, J. R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D. R., et al. (2012). MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* 7:e47656. doi: 10.1371/journal.pone.0047656

Li, H., and Reynolds, J. F. (1995). On definition and quantification of heterogeneity. *Oikos* 73:28.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841.

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–1967. doi: 10.1093/bioinformatics/btp336

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272. doi: 10.1101/gr.097261.109

Li, W., and Godzik, A. (2006). CD-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, W., and Ma, S. Z.-S. (2019). Diversity scaling of human vaginal microbial communities. *Zool. Res.* 40, 587–594. doi: 10.24272/j.issn.2095-8137.2019.068

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 2012, 220–230. doi: 10.1038/nature11550

Ma, Z. S. (1991). Further interpreted Taylor's Power Law and population aggregation critical density. *Trans. Ecol. Soc. China* 1, 284–288.

Ma, Z. S. (2012). Chaotic populations in genetic algorithms. *Appl. Soft Comput.* 12, 2409–2424. doi: 10.1016/j.asoc.2012.03.001

Ma, Z. S. (2013). Stochastic populations, power law and fitness aggregation in genetic algorithms. *Fund. Inform.* 122, 173–206. doi: 10.3233/fi-2013-787

Ma, Z. S. (2015). Power law analysis of the human microbiome. *Mol. Ecol.* 24, 5428–5445. doi: 10.1111/mec.13394

Ma, Z. S. (2018). DAR (diversity–area relationship): Extending classic SAR (species–area relationship) for biodiversity and biogeography. *Ecol. Evol.* 8, 10023–10038. doi: 10.1002/ece3.4425

Ma, Z. S. (2019). A new DTAR (diversity–time–area relationship) model demonstrated with the indoor microbiome. *J. Biogeogr.* 46, 2024–2041. doi: 10.1111/jbi.13636

Ma, Z. S., and Ellison, A. M. (2018). A unified concept of dominance applicable at both community and species scale. *Ecosphere* 9:e02477. doi: 10.1002/ecs2.2477

Ma, Z. S., and Ellison, A. M. (2019). Dominance network analysis provides a new framework for studying the diversity-stability relationship. *Ecol. Monogr.* 89:e01358. doi: 10.1002/ecm.1358

Ma, Z. S., Li, L., and Gotelli, N. J. (2019). Diversity-disease relationships and shared species analyses for human microbiome-associated diseases. *ISME J.* 13, 1911–1919. doi: 10.1038/s41396-019-0395-y

Ma, Z. S., and Li, L. W. (2018). Measuring metagenome diversity and similarity with Hill numbers. *Mol. Ecol. Resour.* 18, 1339–1355. doi: 10.1111/1755-0998.12923

Ma, Z. S., and Li, W. (2019). How and why men and women differ in their microbiomes: medical ecology and network analyses of the microgenderome. *Adv. Sci.* 6:1902054. doi: 10.1002/advs.201902054

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828.

Oh, J., Byrd, A. L., Park, M., Nisc Comparative Sequencing Program, Kong, H. H., Segre, J. A., et al. (2016). Temporal stability of the human skin microbiome. *Cell* 165, 854–866. doi: 10.1016/j.cell.2016.04.008

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60.

Quist, C. W., Gort, G., Mulder, C., Wilbers, R. H. P., Termorshuizen, A. J., Bakker, J., et al. (2017). Feeding preference as a main determinant of microscale patchiness among terrestrial nematodes. *Mol. Ecol.* 17, 1257–1270. doi: 10.1111/1755-0998.12672

Reuman, D. C., Gislason, H., Barnes, C., Mélin, F., and Jennings, S. (2014). The marine diversity spectrum. *J. Anim. Ecol.* 83, 963–979.

Reuman, D. C., Mulder, C., and Cohen, J. (2009). Allometry of body size and abundance in 166 food webs. *Adv. Ecol. Res.* 41, 1–45. doi: 10.1016/S0065-2504(09)00401-2

Reuman, D. C., Zhao, L., and Sheppard, L. W. (2017). Synchrony affects Taylor's law in theory and data. *Proc. Natl. Acad. Sci. U.S.A.* 114:6788.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., et al. (2017). Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat. Methods* 14, 1063–1071.

Shavit, A., Kolumbus, A., and Ellison, A. M. (2016). *Two Roads Diverge in a Wood: Indifference to the Difference Between 'Diversity' and 'Heterogenetiy' Should Be Resisted on Epistemic and Moral Grounds*. Available online at: http://philsci-archive.pitt.edu/12432

Stumpf, M. P. H., and Porter, M. A. (2012). Critical truths about power laws. *Science* 335, 665–666. doi: 10.1126/science.1216142

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature* 189, 732–735. doi: 10.1038/189732a0

Taylor, L. R. (1984). Assessing and interpreting the spatial distributions of insect populations. *Annu. Rev. Entomol.* 29, 321–357. doi: 10.1146/annurev.en.29.010184.001541

Taylor, L. R., Perry, J. N., Woiwod, I. P., and Taylor, R. A. J. (1988). Specificity of the spatial power-law exponent in ecology and agriculture. *Nature* 332, 721–722. doi: 10.1038/332721a0

Taylor, L. R., and Taylor, R. A. J. (1977). Aggregation, migration and population mechanics. *Nature* 265, 415–421. doi: 10.1038/265415a0

Taylor, L. R., Taylor, R. A. J., Woiwod, I. P., and Perry, J. N. (1983). Behavioural dynamics. *Nature* 303, 801–804.

Taylor, L. R., and Woiwod, I. P. (1982). Comparative synoptic dynamics. I. Relationships between inter- and intra-specific spatial and temporal variance/mean population parameters. *J. Anim. Ecol.* 51, 879–906.

Taylor, R. A. J. (1981). The behavioral basis of redistribution. I. the delta-model concept. *J. Anim. Ecol.* 50, 573–586.

Taylor, R. A. J. (2007). Obituary: Roy (L. R.) Taylor (1924–2007). *J. Anim. Ecol.* 76, 630–631. doi: 10.1111/j.1365-2656.2007.01243.x

Tippett, M. K., and Cohen, J. E. (2016). Tornado outbreak variability follows Taylor's power law of fluctuation scaling and increases dramatically with severity. *Nat. Commun.* 7:10668.

Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* 14, 508–522. doi: 10.1038/nrmicro.2016.83

Wearn, O. R., Reuman, D. C., and Ewers, R. M. (2013). Response to comment on Extinction debt and windows of conservation opportunity in the Brazilian Amazon. *Science* 339:271. doi: 10.1126/science.1231618

Xiao, L., Estellé, J., Kiilerich, P., Ramayo-Caldas, Y., Xia, Z., Feng, Q., et al. (2016). A reference gene catalogue of the pig gut microbiome. *Nat. Microbiol.* 1:16161.

Xiao, L., Feng, Q., Liang, S., Sonne, S. B., Xia, Z., Qiu, X., et al. (2015). A catalog of the mouse gut metagenome. *Nat. Biotechnol.* 33, 1103–1108.

Zhang, Z., Geng, J., Tang, X., Ma, Z. S., and Shi, P. (2014). Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *ISME J.* 8:881. doi: 10.1038/ismej.2013.185

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 38:e132. doi: 10.1093/nar/gkq275

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 June 2020
doi: 10.3389/fmicb.2020.01121

Exploring the Bacterial Impact on Cholesterol Cycle: A Numerical Study

Mélanie Bourgin[1], Simon Labarthe[2], Aicha Kriaa[1], Marie Lhomme[3,4], Philippe Gérard[1], Philippe Lesnik[3], Béatrice Laroche[2], Emmanuelle Maguin[1] and Moez Rhimi[1]**

[1] Micalis Institute, Université Paris-Saclay, INRAE, AgroParisTech, Jouy-en-Josas, France, [2] Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France, [3] INSERM, UMRS 1166, Sorbonne Universités, Hôpital Pitié-Salpétrière, Paris, France, [4] ICANalytics, Institute of Cardiometabolism and Nutrition (IHU-ICAN, ANR-10-IAHU-05), Paris, France

High blood cholesterol levels are often associated with cardiovascular diseases. Therapeutic strategies, targeting different functions involved in cholesterol transport or synthesis, were developed to control cholesterolemia in human. However, the gut microbiota is also involved in cholesterol regulation by direct biotransformation of luminal cholesterol or conversion of bile salts, opening the way to the design of new strategies to manage cholesterol level. In this report, we developed for the first time a whole-body human model of cholesterol metabolism including the gut microbiota in order to investigate the relative impact of host and microbial pathways. We first used an animal model to investigate the ingested cholesterol distribution *in vivo*. Then, using *in vitro* bacterial growth experiments and metabolite measurements, we modeled the population dynamics of bacterial strains in the presence of cholesterol or bile salts, together with their bioconversion function. Next, after correct rescaling to mimic the activity of a complex microbiota, we developed a whole body model of cholesterol metabolism integrating host and microbiota mechanisms. This global model was validated with the animal experiments. Finally, the model was numerically explored to give a further insight into the different flux involved in cholesterol turn-over. According to this model, bacterial pathways appear as an important driver of cholesterol regulation, reinforcing the need for development of novel "bacteria-based" strategies for cholesterol management.

Keywords: microbiota, holobiont, microbiome, functional ecology, cholesterol metabolism, whole body model, mathematical model, system biology

OPEN ACCESS

Edited by:
Hyun-Seob Song, University of Nebraska-Lincoln, United States

Reviewed by:
Mark Mc Auley, University of Chester, United Kingdom
Paritosh Pande, Pacific Northwest National Laboratory (DOE), United States

***Correspondence:**
Simon Labarthe
simon.labarthe@inrae.fr
Moez Rhimi
moez.rhimi@inrae.fr

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *10 January 2020*
Accepted: *05 May 2020*
Published: *10 June 2020*

Citation:
Bourgin M, Labarthe S, Kriaa A, Lhomme M, Gérard P, Lesnik P, Laroche B, Maguin E and Rhimi M (2020) Exploring the Bacterial Impact on Cholesterol Cycle: A Numerical Study. Front. Microbiol. 11:1121. doi: 10.3389/fmicb.2020.01121

1. INTRODUCTION

Cholesterol plays an essential role in the human body (Arnold and Kwiterovich, 2003). It is a key component of cellular membranes, being involved in membrane fluidity, cellular organization, and signaling (Ikonen, 2008; Mesmin and Maxfield, 2009). Cholesterol also serves as a precursor of many biological molecules including bile acids, oxysterols, steroid hormones, and vitamin D (Schroepfer Jr, 2000; Tabas, 2002). In humans, 30% of total body cholesterol derive from the diet (exogenous or dietary cholesterol), the remaining 70% are mainly synthesized in the liver (endogenous cholesterol) (Gylling, 2004). Over the last decades, several studies have aimed at deciphering the pathways involved in cholesterol homeostasis (Gylling, 2004; Iqbal and Hussain, 2009; Russell, 2009; Millar and Cuchel, 2018). In mammalian bodies, cholesterol balance is maintained by tightly regulated interactions between cholesterol synthesis, bile salts (BS) synthesis, absorption, and excretion.

Although cholesterol exhibits multiple physiological functions, high blood cholesterol levels are often associated to cardiovascular diseases (CVD), the leading cause of death in the world (World Health Organization, 2017). Current therapeutic strategies mainly target host cholesterol biosynthesis or transport, with no cholesterolemia reduction in a significant proportion of patients and major side effects (Thompson et al., 2002; Potiron et al., 2015). Recently, the gut microbiota has emerged as a key player that influences metabolic health and disease (Doré et al., 2017). It is possible that the gut microbiota could contribute to cholesterol metabolism mostly through (i) bacterial deconjugation of BS by bile salts hydrolase (BSH) enzymes and (ii) cholesterol conversion into coprostanol, a non-absorbable molecule excreted in feces (Begley et al., 2006; Gérard et al., 2007).

Accumulating data regarding each pathway have been reported (Swann et al., 2011; Jones et al., 2012; Joyce et al., 2014; Ridlon et al., 2014; Kriaa et al., 2019), but functional and mechanistic insights into their impact on whole-body cholesterol homeostasis are still lacking. To better understand the complex interplay between each human compartments, whole-body mathematical models were previously described (van de Pas et al., 2010, 2012; Mc Auley et al., 2012; Morgan et al., 2016; Read and Holmes, 2017). However, existing models were focused on human cholesterol biosynthesis or lipoprotein metabolism and do not include the gut microbiota as a crucial and new player in this complex multicompartments cycle (Pool et al., 2018).

The aim of this work is to provide an estimation of the impact of the microbial activity on the cholesterol cycle. Since BS are naturally present in the small intestine lumen where sterol and BS absorption take place, we hypothesized that bacterial BS deconjugation and cholesterol-to-coprostanol conversion could impact the cholesterol fate in the host body compartments. In order to assess this impact, we adopted an integrative approach in which literature based knowledge as well as *in vitro* and *in vivo* experimental data are used to generate a whole-body mathematical model of cholesterol metabolism in human holobiont including its associated gut microbiota. In a dedicated experiment, cholesterol was tracked in mice, in order to investigate the distribution of ingested cholesterol in different host compartments, and determine the amount of bioavailable cholesterol in the gastrointestinal tract. We also characterized *in vitro* BS deconjugation and cholesterol-to-coprostanol conversion activity in several commensal bacterial strains. Finally, we developed a mathematical model to link all the experimental data, starting from existing models in the literature (van de Pas et al., 2010, 2012; Mc Auley et al., 2012; Morgan et al., 2016). The different bacterial pathways for cholesterol and BS metabolism were calibrated and integrated in the model, allowing for differential comparison. Numerical exploration was then conducted to decipher the relative impact of the host and the microbiota metabolisms on the overall cholesterol cycle.

2. METHODS AND MATERIALS

2.1. Chemicals, Media, and Reagents

Deuterated cholesterol-d5 [cholesterol-2,2,4,4,6-d5] was purchased from Medical Isotopes, Inc. Medium-chain triglycerides (MCT) were purchased from Now food (Healthcenter). Reagents and standards were supplied for sterol extraction by gas-chromatographic/mass-spectrometry analysis (GC/MS). Chloroform, cyclohexane, methanol were purchased from Merck. Butylated Hydroxytoluene (BHT) was supplied from Sigma and used to prepare BHT solution in methanol (5 $mg\,mL^{-1}$). Hexandiethylether was purchased from VWR Chemicals. Analytical standard of desmosterol-d6 [cholest-5,24-dien-3-ol] was purchased (Avanti® Polar lipid, Inc.). Desmosterol solution was prepared (200 $\mu mol\,L^{-1}$) with chloroform and used for cholesterol quantification. Derivatization reagent N,O-bis(trimethylsilyl)trifluoroacetamid (BSTFA) with 1% trimethylchlorosilane (TMCS) was obtained from REGIS technology, Inc. Cholesterol, sodium taurocholate hydrate, sodium glycocholate hydrate, sodium taurodeoxycholic acid hydrate and sodium glycodeoxycholic acid hydrate, sodium taurochenodeoxycholic acid hydrate and sodium glycochenodeoxycholic acid hydrate, ninhydrin, and trichloroacetic acid were purchased from Sigma-Aldrich. Bacteria were grown in Brain Heart Infusion-Yeast extract-Hemin medium (BHI-YH) containing: 5 $g\,L^{-1}$ of yeast extract, 5 $mg\,L^{-1}$ of hemin, 2 $mg\,L^{-1}$ of vitamin K, and 0.5 $g\,L^{-1}$ of cysteine (all products from Sigma-Aldrich). This media was supplemented when necessary with cholesterol and BS according to the supplier recommendations.

2.2. Bacterial Growth Procedure and BSH Assays

Bacteroides xylanisolvens XB1A and *Bacteroides* sp. D8 were grown in standard BHI-YH broth. All cultures were grown at 37°C in anaerobic conditions (Freter chamber Jacomex, France, 85% N2, 10% H2, 5% CO2) during 24 h. Effect of bile acids on bacterial growth was tested in BHI-YH supplemented with 1 and 30 mM of bile acids (Sigma). For cell lysate preparation pellets were washed twice in 100 mM sodium-phosphate buffer pH 6.5 and resuspended in the same buffer. Cell disruption was done by sonication at 4°C during 1 min (three cycles of 10 s pulses at amplitude of 40%) using a Vibra-Cell TM 72408 Sonicator then, cell debris were removed by centrifugation (12,000 g, 30 min at 4°C). Protein concentration was determined by measuring the UV absorption at 280 nm using a Nanodrop device (Thermo Fisher Scientific). The BSH activity was measured by the determination of the amount of the released amino acid residues using two BS as previously reported (Tanaka et al., 1999). At standard conditions, the reaction mixture contained 50 μL of enzyme preparation at a suitable dilution, 10 mM glyco and tauro-conjugated BS with 10 mM sodium phosphate pH 6.5 in a final volume of 1 mL.

The mixture was incubated for 30 min at 37°C then the reaction was stopped by adding 20% trichloroacetic acid and incubated at 37°C during 30 min. Subsequently, the reaction mixture was centrifuged (12,000 g, 15 min, 4°C) and the supernatant was recovered. For 200 μL of sample we added 500 μL of 1% ninhydrin, 1.2 mL of glycerol 30% and 200 μL of 500 mM citrate buffer pH 5.5. Then, the amount of amino acid released from conjugated bile acids was determined by measuring the absorbance at 570 nm using a UV-spectrophotometer

(Spectro-biochrom LibraS11). One unit of BSH activity was defined as the amount of enzyme catalyzing the release of 1 μmol of amino acids per min under the above specified conditions.

2.3. Animals and Experimental Design

Eleven-week-old male wild type C57BL/6 mice were purchased from the Laboratory Janvier (Le Gesnest, St Isle, France), and maintained in our animal facilities (INRA, UMR1319 Micalis, Anaxem facilities) under specific pathogen-free conditions. Throughout the experimental period, mice were provided free access to water and a standard diet containing 0.02% of cholesterol (SAFE, R03-40) (Wang and Carey, 2003). To curtail coprophagy during the study, animals were housed in individual metabolic cages with wire mesh bottoms (Wang et al., 2001). All procedures were performed according to the European Community Rules and approved by the Animal Care Committee (C2E-45 COMETHEA) with authorization number A78-322-6. Then, a group of experimental mice received an oral dose of 0.6 mg deuterated cholesterol-d5 dissolved in 200 μL MCT ($n = 6$) and a group of control mice received 200 μL of MCT as previously reported (Jakulj et al., 2016). After 3 days, feces were recovered for sterol quantification ($n = 3$). Blood collected and tissue samples were collected following animal euthanasia. Serum was collected after centrifugation (3,000 g during 10 min, 4°C) in presence of 2 mM EDTA. All samples were frozen in liquid nitrogen then stored at −80°C.

2.4. Sterol Extraction and Quantification

Plasma, feces and tissue sterols were extracted in the presence of an internal standard, deuterated desmosterol-d6 (200 μmol L^{-1}) according to the Folch method with some modifications (Folch et al., 1957). Each tissue and feces was dried (approximately 0.3 g), powdered and homogenized in chloroform-methanol (2:1 v/v) at 63°C overnight (Igel et al., 2003). The same protocol was used for plasma aliquots (200 μL) after previous homogenization during 1 h. After addition of water (1:1 v/v), samples were centrifuged and the organic phase was collected. The organic dried extract, was resuspended in 2 mL methanol-NaOH 1 M, 40 μL BHT-Methanol and 40 μL methanol-EDTA at 60°C during 1 h allowing the lipids saponification. Subsequently, lipids were again extracted using hexan-diethyl-ether (1:1 v/v). After mixture and centrifugation of samples, the organic phase was collected and dried followed by reconstitution in 1.4 mL of cyclohexane. The silylation of sterols was performed with 60 μL of BSTFA with 1% TMCS and 1 h incubation at 60°C. After homogenization and centrifugation pellets were suspended in 60 μL of cyclohexane. The samples were stored at −80°C until the GC/MS analysis.

2.5. Mathematical Model of Specialized Bacterial Strains in Cholesterol and BS Metabolisms

Dynamical systems describing bacterial growth and metabolite concentration dynamics were fitted with the growth assays of *Bacteroides xylanisolvens* XB1A and *Bacteroides* sp D8. A minimal logistic ordinary differential equation (ODE) (resp. delayed differential equation (DDE)) was designed to model *Bacteroides* sp. D8 (resp. *Bacteroides xylanisolvens* XB1A) growth, supplemented by metabolic and repression mechanisms (see Results section for the detailed models). The equation parameters were inferred with a Bayesian inference method based on the DRAM sampling method (Haario et al., 2006) and a normal likelihood function, or linear regression (for BSH assays) after removal of outliers. Markov chains convergence was checked with the Geweke criterion.

TABLE 1 | MCMC parameter estimation results.

Parameter	Mean	Std	Geweke
Bacteroides Sp D8			
$\mu_{B_{spD8}}$	0.44772	0.0281	0.98987
k_{ccD8}	0.27441	0.21987	0.76473
K_{D8}	1.6681	1.2765	0.94418
Bacteroides xylanosolvens			
μ_{Bxyl}	1.9375	0.95771	0.82175
β_{Bxyl}	1.1186	0.57418	0.8144
δ	24.495	0.29425	0.99817
K_{Bxyl}	0.10439	0.0936	0.88024

We indicate, for each parameter, the mean and the standard deviation of the posterior parameter distribution given by the MCMC Bayesian estimation, together with the Geweke index of the corresponding Markov chains. The corresponding posteriors are given in ***Figure S1****.*

2.6. Whole Body Model of Cholesterol Metabolism

We built our compartment dynamic model on the global structure of a previously reported whole-body model (Mc Auley et al., 2012; Morgan et al., 2016), which included the enterohepathic BS cycle, the plasmatic regulation and transport of cholesterol from the intestine toward the peripheral tissues and the liver, the coupling between bile acids and cholesterol metabolism through bile production, and the intestinal flux: dietary influx, hepatic cholesterol release in the digestive track, and excretion in feces. As in Morgan et al. (2016), a luminal compartment was introduced including the luminal primary BS and the luminal cholesterol, to which we added the microbiota. Furthermore, we simplified several uptake and transport processes that were not relevant for our study, following (van de Pas et al., 2011, 2012). A global view of the model is presented in **Figure 3**, the precise model description can be found in section 3.3, and the model parameter in the **Table S1**, **Table 1**).

2.7. Whole-Body Model Calibration

We adapted a strategy previously used for model calibration (van de Pas et al., 2011, 2012). Documented steady-state flux and levels of cholesterol in mice were collected, discarding at this stage the bacterial metabolism. The unknown flux were reconstructed through mass-conservation equations: at steady state, flux balance equations involving the unknown flux are derived. Additional equations are set to conserve the ratio of transport flux between blood and liver compartments. At end, as many conservation equations as unknown flux are defined. All

the parameters were then obtained straightforwardly by direct computation of the parameters, given the flux and cholesterol levels at steady-state, as indicated in **Table S1**. Next, we upscaled the growth models of specialized strains obtained *in vitro* to mimic the metabolism of a complex microbiota *in vivo*: the dynamics of coprostanol degradation was calibrated on *in vivo* data collected from the literature (see **Table S1** and **Table 2** for references), and the BS degradation was deduced from the BSH activity measured during the animal experiments. Finally, time was rescalled between the *in vitro* and the whole-body model, which allowed to replace the DDE for *Bacteroides xylanisolvens* XB1A by a non-delayed ODE (cf section 3.3 for details).

2.8. Numerical Implementation

The model was implemented in the Matlab software (MathWorks, Natick, MA, USA). The time integration of the ODEs and DDEs was achieved with, respectively, the ode15s and the dde23 matlab functions. Bayesian inference was performed with the MCMC matlab toolbox (https://mjlaine.github.io/mcmcstat/) (Haario et al., 2001, 2006). Linear regression was performed with the R *lm* function.

2.9. Sensitivity Analysis

We first studied the local sensitivity of model outputs respectively to the bacterial levels. Namely, we applied to the BS and cholesterol converter carrying capacity (respectively $PBSD_{MAX}$ and CCC_{MAX}) a multiplicative coefficient $q \in [\frac{1}{100}, \frac{1}{50}, \frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{1}{2}, 2, 5, 10, 20, 50, 100]$, and we observed the impact of these variations on steady-state cholesterol and BS flux and levels. We then studied the global sensitivity of our model to flux parameters by computing parameter Sobol index (Saltelli et al., 1999) and Partial Correlation Coefficients (PCC) (Saltelli et al., 2000) of the $d = 14$ main parameters involved in the flux of the BS and cholesterol cycles. Namely, we selected for the BS enterohepathic cycle the bacterial carrying capacity of BS converters ($PBSD_{MAX}$), BS synthesis rate (k_{HBSs}), BS release in the lumen (k_{HBSo}) and absorption by the intestinal epithelium (k_{LPBSa}). For the cholesterol cycle, we selected the bacterial capacity for cholesterol converters (CCC_{MAX}), cholesterol synthesis rates (ICS_{max}, HCS_{max} and PCS_{max} for respectively the intestinal epithelium, the liver and the peripheral tissues, that were shifted all together), transport from blood to liver ($k_{LDL,ha}$ and $k_{HDL,ha}$, shifted conjointly), transport from liver to blood (k_{HCo}), cholesterol release (BCR_{max}) and dietary intake (f_{meal}). We sampled uniformly ($n = 11 \cdot 10^5$ samples) the parameter hypercube ranging in $\pm$ 50 % the basal value obtained after model calibration, except for the bacterial carrying capacities that were uniformly shifted between 0.01 and 100 times the basal value, with the fast99 method (Saltelli et al., 1999). The R package sensitivity (https://cran.r-project.org/web/packages/sensitivity Iooss and Lemaître, 2015) was used to build the experimental design and to compute the first order Sobol index and the PCC with the function *fast99* and *pcc* respectively.

3. RESULTS

3.1. *In-vivo* Cholesterol Body Distribution

To check the cholesterol body distribution, we gave to three mice a standard diet supplemented with a dose of deuterated cholesterol. The distribution of labeled cholesterol among compartments after 3 days is displayed in **Figure 1**. We observed that about half of the labeled cholesterol was excreted in the feces (48.1%) and about one quarter (26.4%) was stocked in the mice tissues (plasma, peripheral tissue and liver) while the last quarter (25.5%) was still circulating in the intestinal lumen and tissues. This indicates that the cholesterol pool available for bacterial biotransformation represents an important fraction of the ingested cholesterol, suggesting that bacteria could have a noticeable impact on cholesterol fate.

3.2. *In vitro* Data-Based Models of Bacterial Cholesterol and BS Metabolism

We next used the bacterial growth assays to model the bacterial population dynamics and their functions related to cholesterol and BS. For each assay, we tested several models and chose the simplest one, i.e., the model providing the best trade off between goodness of fit and number of parameters.

3.2.1. *Bacteroides* sp. D8 Cholesterol Conversion

We first modeled the dynamics of *Bacteroides* sp D8 normalized density ($B_{spD8} := \frac{[B_{spD8}]}{[B_{spD8}]_{max}}$, where $[B_{spD8}]$ is the bacterial concentration ($[\mathrm{CFU\,mL^{-1}}]$) and $[B_{spD8}]_{max}$ is the maximal observed bacterial concentration), with the logistic equation

$$\partial_t B_{spD8} = \mu_{B_{spD8}} B_{spD8} \left(1 - B_{spD8}\right). \tag{1}$$

Note that no dependency with cholesterol levels was introduced in the logistic model. This simple model has been selected because we aimed at modeling the bacterial growth in a complex nutritional environment, and not only the catabolic capabilities obtained from cholesterol degradation. The multiple pathways activated during the growth on BHI-YH are summed up in the growth rate of the logistic model.

Cholesterol (Cl) is converted to coprostanol (Cp) so that their respective fraction follow equations

$$\partial_t Cl = -k_{ccD8} \frac{B_{spD8} Cl}{K_{D8} + B_{spD8}}, \qquad \partial_t Cp = k_{ccD8} \frac{B_{spD8} Cl}{K_{D8} + B_{spD8}}. \tag{2}$$

The parameter $\mu_{B_{spD8}}$, k_{ccD8}, and K_{D8} were inferred with Bayesian inference, processing conjointly the growth assays with different initial BS concentrations. We used the uniform prior $\mu_{B_{spD8}} \sim U_{(0.1\alpha_{B_{spD8}}, 2\alpha_{B_{spD8}})}$, $k_{ccD8} \sim U_{(10^{-5},1)}$, and $K_{D8} \sim U_{(10^{-3},4)}$ where $\alpha_{B_{spD8}}$ is an approximation of the B_{spD8} growth rate during the log-phase. The posterior parameter distributions are displayed in **Figure S1** and the mean and variance values can be found in **Table 1**, together with the corresponding geweke index of markov chain convergence. Bacteria and metabolite levels and data fit are displayed in **Figure 2**.

TABLE 2 | Parameters used for the calibration of the whole-body cholesterol cycle.

Parameter	Value	Unit	Description	References
Cholesterol steady state fluxes in the whole body model				
$ss_{k_{in}}$	0.78	mg day^{-1}	Steady state dietary cholesterol influx.	van de Pas et al., 2011
ss^{ref}_{kLCe}	0.8734	mg day^{-1}	Reference total steady state fecal cholesterol excretion.	Van der Velde et al., 2007
$ss^{ref}_{chol,copro}$	0.1	mg day^{-1}	Steady state excreted coprostanol to cholesterol ratio.	Sekimoto et al., 1983
ss_{kLCe}	1.2352	mg day^{-1}	Total steady state fecal cholesterol excretion. $ss_{kLCe} = (1 - ss^{ref}_{chol,copro}/(1 + ss^{ref}_{chol,copro}))ss^{ref}_{kLCe}$	*MC*
ss_{kCC}	0.12352	mg day^{-1}	Steady state conversion of cholesterol to coprostanol. $ss_{kCC} = ss^{ref}_{chol,copro}/(1 + ss^{ref}_{chol,copro})ss^{ref}_{kLCe}$	*MC*
ss_{kLCo}	0.4852	mg day^{-1}	Steady state direct luminal release of intestinal cholesterol.	Van der Velde et al., 2007
ss_{BCRmax}	0.1941	mg day^{-1}	Steady state hepatic cholesterol biosynthesis.	van de Pas et al., 2011
ss_{kLCa}	0.097	mg day^{-1}	Steady state uptake of luminal cholesterol $ss_{kLCa} = ss_{k_in} + ss_{BCRmax} + ss_{kLCo} - ss_{kLCe} - ss_{kCC}$.	*MC*
$ss_{ICS_{max}}$	0.87	mg day^{-1}	Steady state intestinal cholesterol biosynthesis.	van de Pas et al., 2011
$ss_{1_{\theta_l},kICo}$	0.097	mg day^{-1}	Steady state uptake of intestinal cholesterol by HDL.	Van der Velde et al., 2007
$ss_{\theta_l,kICo}$	0.3882	mg day^{-1}	Steady state uptake of intestinal cholesterol by LDL $ss_{\theta_l,kICo} = ss_{kLCa} + ss_{ICS_{max}} - ss_{1,\theta_l,kICo} - ss_{kLCo}$.	*MC*
$ss_{HCS_{max}}$	1.75	mg day^{-1}	Steady state hepatic cholesterol biosynthesis.	van de Pas et al., 2011
ss_{kHCest}	0.9705	mg day^{-1}	Steady state hepatic cholesterol esterification rate.	Van der Velde et al., 2007
$ss_{kHCunest}$	0.9705	mg day^{-1}	Steady state rate of unesterification $ss_{kHCunest} = ss_{kHCest}$.	*MC*
$ss_{\theta,kHCo}$	0.9705	mg day^{-1}	Steady state hepatic cholesterol uptake by LDL	van de Pas et al., 2011
$ss^{ref}_{1,\theta_H,kHCo}$	0.7764	mg day^{-1}	Reference Steady state hepatic cholesterol uptake by HDL	Van der Velde et al., 2007
ss^{ref}_{kLDLha}	1.1646	mg day^{-1}	Reference steady state absorption of LDL cholesterol by liver.	Van der Velde et al., 2007
ss^{ref}_{kHDLha}	1.7469	mg day^{-1}	Reference steady state HDL cholesterol absorption by liver	van de Pas et al., 2011
ss_{kLDLha}	1.2542	mg day^{-1}	Steady state absorption of LDL cholesterol by the liver. $ss_{kLDLha} = (ss_{\theta_l,kICo} + ss_{\theta_H,kHCo})/(1 + \frac{ss^{ref}_{kLDLpa}}{ss^{ref}_{kLDLha}})$.	*MC*
ss_{kHDLha}	1.5856	mg day^{-1}	Steady state absorption of LDL cholesterol by the liver. $ss_{kHDLha} = (ss_{HCS_{max}} + ss_{kLDLha} - ss_{\theta_H,kHCo} - ss_{kHBSs} - ss_{BCRmax})/(\frac{ss^{ref}_{1,\theta_H,kHCo}}{ss^{ref}_{kHDLha}} - 1)$.	*MC*
$ss_{1,\theta_H,kHCo}$	0.7047	mg day^{-1}	Steady state uptake of hepatic cholesterol by HDL $ss_{1,\theta_H,kHCo} = (ss_{HCS_{max}} + ss_{kLDLha} - ss_{\theta_H,kHCo} - ss_{kHBSs} - ss_{BCRmax})/(1 - \frac{ss^{ref}_{kHDLha}}{ss^{ref}_{1,\theta_H,kHCo}})$.	*MC*
ss_{B_H}	2.9115	mg day^{-1}	Total steady state absorption of cholesterol by the liver from the blood. $ss_{B_H} = ss_{kLDLha} + ss_{kHDLha}$	*MC*
$ss_{PCS_{max}}$	1.16	mg day^{-1}	Steady state peripheral cholesterol biosynthesis.	van de Pas et al., 2011
ss^{ref}_{kLDLpa}	0.0970	mg day^{-1}	Ref. steady state absorption of LDL cholesterol by peripheral tissues.	Van der Velde et al., 2007
ss_{kLDLpa}	0.1045	mg day^{-1}	Steady state absorption of LDL cholesterol by the peripheral tissues. $ss_{kLDLpa} = (ss_{\theta_l,kICo} + ss_{\theta_H,kHCo})/(1 + \frac{ss^{ref}_{kLDLha}}{ss^{ref}_{kLDLpa}})$.	*MC*
$ss_{1,\theta_P,kPCo}$	0.7839	mg day^{-1}	Steady state uptake of peripheral cholesterol by HDL $ss_{1,\theta_P,kPCo} = ss_{PCS_{max}} + ss_{kLDLpa} - ss_{kPloss}$.	*MC*
ss_{kPloss}	0.4852	mg day^{-1}	Steady state cholesterol loss by peripheral metabolism $ss_{kPloss} = ss_{kLCa} + ss_{HCS_{max}} + ss_{PCS_{max}} + ss_{ICS_{max}} - ss_{BCRmax} - ss_{kHBSs} - ss_{kLCo}$.	*MC*

We define for each compartment, the steady state fluxes involved in the cholesterol transport processes and a reference in the literature. MC: parameter derived from mass conservation arguments with the given equation. BS cycle steady state fluxes are given in ***Table S6*** *(Supplementary Material).*

3.2.2. *Bacteroides xylanosolvens* BS Conversion

We then modeled the *Bacteroides xylanosolvens* normalized population dynamics ($Bxyl := \frac{[Bxyl]}{[\bar{Bxyl}]_{max}}$, where $[Bxyl]$ is the bacterial concentration ($[\mathrm{CFU\,mL^{-1}}]$) and ($[\bar{Bxyl}]_{max}$) is the maximal observed bacterial concentration), with a logistic equation and a repression term that model the bacterial sensitivity to the primary bile salts (PBS) with delay δ:

$$\begin{aligned}\partial_t Bxyl(t) &= \mu_{Bxyl} Bxyl(t)\left(1 - Bxyl(t)\right) \\ &- \beta_{Bxyl}\frac{Bxyl\,[PBS](t-\delta)}{K_{Bxyl} + [PBS](t-\delta)}.\end{aligned} \tag{3}$$

The deconjugation of PBS into secondary bile salts (SBS) follows the equations

$$\begin{aligned}\partial_t [PBS] &= -\tilde{k}_{Bxyl}(Bxyl)[PBS](t), \\ \partial_t [SBS] &= \tilde{k}_{Bxyl}(Bxyl)[PBS](t).\end{aligned} \tag{4}$$

The parameter $\tilde{k}_{Bxyl}(Bxyl)([\mathrm{h^{-1}}])$ representing the degradation rate induced by the bacteria (that varies with *Bxyl*) was given by the enzyme assays with the following heuristic.

The enzyme assays allowed to measure $A_{BSH}[\mathrm{nmol\,min^{-1}\,mg_{prot}^{-1}}]$ which was the SBS production rate by gram of total proteins in the sample for an initial BS concentration $[BS]_0([nmol.mL^{-1}])$ in the growth media. Note that A_{BSH} varied with *Bxyl* so that $A_{BSH} := A_{BSH}(Bxyl)$. Hence

$$\tilde{k}_{Bxyl}(Bxyl) := \tilde{T}_k \frac{\lambda(Bxyl)}{[BS]_0} A_{BSH}(Bxyl)$$

where $\tilde{T}_k = 60\ \mathrm{min\,h^{-1}}$ was a time rescaling coefficient and $\lambda(Bxyl)([mg_{prot}.mL^{-1}])$ was the total protein production by *mL* of the population *Bxyl*. The dependence of BSH activity $A_{BSH}(Bxyl)$ to bacteria levels was first approximated by linear regression on the data, giving $A_{BSH}(Bxyl) := a_{BSH}Bxyl + b_{BSH}$ with $a_{BSH} = 19.2466$ ($p < 2.10^{-3}$) and $b_{BSH} = -0.9437$ ($p = 0.807$). As the intercept value was not significant, b_{BSH} was left null, so that $A_{BSH}(Bxyl) := a_{BSH}Bxyl$. The total protein levels in bacterial cells $\lambda(Bxyl)$ was derived from the literature by writing

$$\lambda(Bxyl) := \tilde{C}_\lambda d_c (1 - c_w) c_p V_c [\bar{Bxyl}]_{max} Bxyl$$

with $\tilde{C}_\lambda = 10^{-9}\mathrm{mg\,g^{-1}mL\,\mu m^{-3}}$ a concentration rescaling coefficient, $d_c(\mathrm{g\,mL^{-1}})$ the bacterial mass density, $c_w([-])$ the proportion of water in the cell, $c_p([-])$ the fraction of protein in the dry mass, and V_c the volume of one bacteria, assumed to be $1\mu m^3.CFU^{-1}$. The value of the different parameters can be found in **Table S2**.

Hence, noting $k_{Bxyl} := \frac{\tilde{T}_k a_{BSH}}{[BS]_0}\tilde{C}_\lambda d_c (1 - c_w) c_p V_c [\bar{Bxyl}]_{max}$, we rewrite Equation (4) with

$$\begin{aligned}\partial_t [PBS] &= -k_{Bxyl} Bxyl(t)^2 [PBS](t), \\ \partial_t [SBS] &= k_{Bxyl} Bxyl(t)^2 [PBS](t).\end{aligned} \tag{5}$$

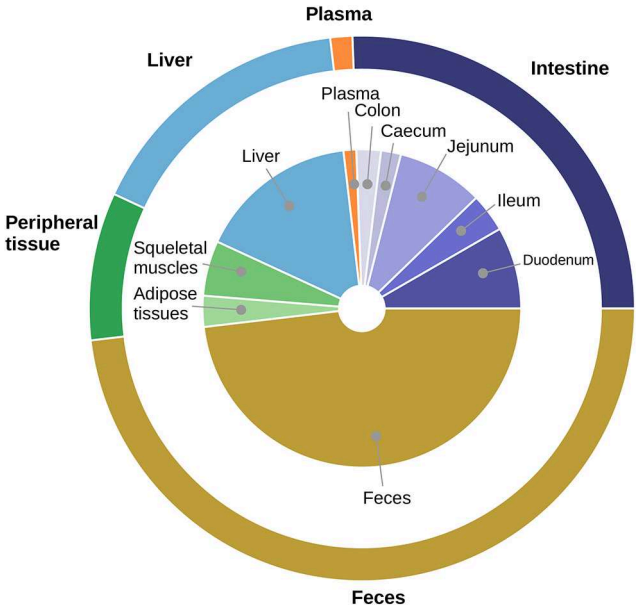


FIGURE 1 | Averaged distribution of labeled cholesterol in mice. The proportion of D5 labeled cholesterol in each compartment 3 days after ingestion is displayed. We obtained the average amount ($n = 3$) of cholesterol in each compartment by GC/MS with internal standard (see section Materials and Methods). During experiments, cholesterol distribution was measured with a finer granularity than in the mathematical model: the central pie chart represents the distribution among the different compartments measured experimentally whereas the external pie chart indicates the corresponding distribution compartments represented in the mathematical model. The external pie is obtained by pooling the corresponding sub-compartments sampled during experiments. We observed that half of the labeled cholesterol ended up in the feces, while about one quarter remained in the intestinal compartment.

The parameters μ_{Bxyl}, β_{Bxyl}, δ and K_{Bxyl} were inferred with the uniform prior $\mu_{Bxyl} \sim U_{(0.6\alpha_{Bxyl},10\alpha_{Bxyl})}$, $\beta_{Bxyl} \sim U_{(10^{-4},8)}$, $K_{Bxyl} \sim U_{(10^{-3},3)}$, and $\delta \sim U_{(15,25)}$ where α_{Bxyl} approximates the Bxyl growth rate during the log-phase from the data. The posterior distributions are displayed in **Figure S1** and the mean and variance values can be found in **Table 1**, together with the corresponding geweke index of markov chain convergence. Model output and data fit are displayed in **Figure 2**.

3.3. Whole Body Model Including the Gut Microbiota

We first detailed the luminal intestinal compartment, where the bacterial activity takes place: we upscaled the *in vitro* model to be representative of bacterial activities observed *in vivo*. We next presented the remaining processes of the whole body model of cholesterol cycle, all located in host compartments. A global view of the model is presented in **Figure 3**. A nomenclature of the different unknowns of the model can be found in **Table S3**.

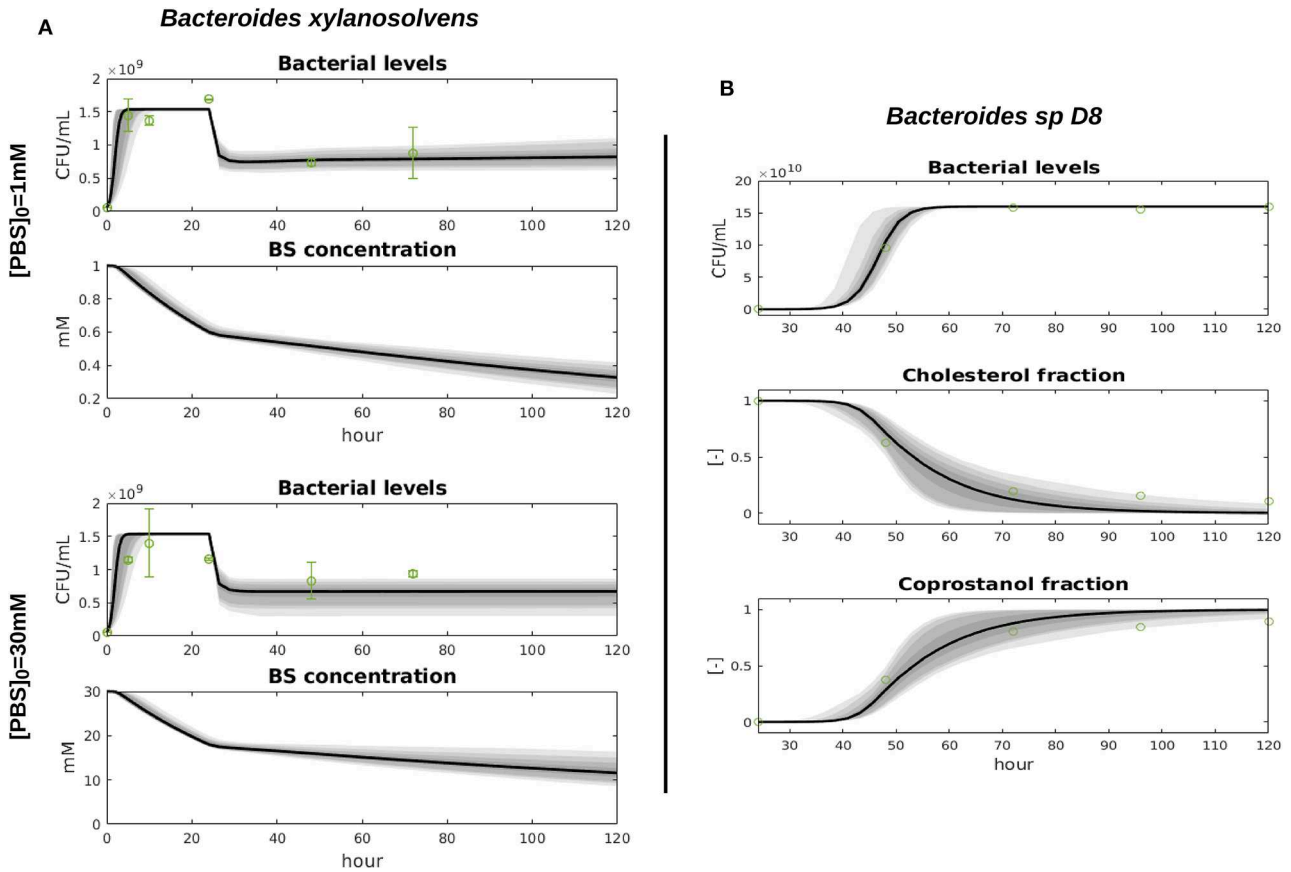


FIGURE 2 | Fit of the bacterial growth models with the data. We display the predictive envelopes of the model by sampling parameter values from the posterior distributions: the black bold line represent the median simulation. The gray areas in the plot correspond to 50, 90, 95, and 99% posterior regions. Data mean and 95% confidence intervals are plotted with green dots and error bars.

3.3.1. Luminal Compartment Including Microbiota

Bacterial growth: the dynamics of the functional bacterial populations involved in cholesterol-to-coprostanol conversion (CCC [−]) or primary-bile-salts deconjugation ($PBSD$ [−]) in the gut were derived from the *in vitro* experiments by taking

$$\partial_t CCC = \mu_{CCC} CCC(CCC_{MAX} - CCC), \tag{6}$$

$$\partial_t PBSD = \mu_{PBSD} PBSD(PBSD_{MAX} - PBSD) - d_{PBSD} \frac{[LPBS]PBSD}{(K_{PBSD} + [LPBS])}. \tag{7}$$

The rescaled growth rates $\mu_{CCC} \coloneqq 24\mu_{B_{spD8}} \frac{b_{gut,max}}{[B_{spD8}]_{max}}$ (day^{-1}) and $\mu_{PBSD} \coloneqq 24\mu_{Bxyl} \frac{b_{gut,max}}{[Bxyl]_{max}}$ (day^{-1}) were derived from the inferred growth rates of Equations (1)–(3), and $b_{gut,max} = 5.0 * 1e9$ CFU mL^{-1}, the bacterial levels in the small intestine (Bazett et al., 2016). The terms $d_{PBSD} \coloneqq 24\beta_{Bxyl}$ and $K_{PBSD} \coloneqq w_{PBS} K_{Bxyl}$ set the $PBSD$ population susceptibility to luminal PBS concentration $[LPBS]$(mg.L^{-1}), where $w_{PBS} \coloneqq 467,847$mg.mmol^{-1} was the molecular weight of PBS. Note that we removed in (7) the delay term δ of Equation (3). Indeed, after time rescaling, the delay had very little impact: when we replaced Equation (7) by its time-delayed original version (3), we observed a relative difference lower than 10^{-6} in $L^2(0,T)$ norm. The parameters CCC_{MAX} and $PBSD_{MAX}$ represent the bacterial carrying capacity. They are set to 1 in the basal simulations but will be shifted during model exploration (cf. sections 3.5, 3.6).

Luminal primary bile salts (LPBS) dynamics: next, we adapted the *in vitro* BS conversion model to the BSH activity of a complex microbial *in vivo* with a suitable upscale of the parameters. Namely, k_{PBSD}, the rate of primary to secondary BS conversion by the microbiota, was derived from the formula

$$k_{PBSD} \coloneqq k_{Bxyl} \frac{A_{BSH,mic} b_{gut,max}}{A_{BSH}([Bxyl]_{max})}$$

where $A_{BSH,mic}$([nmol min^{-1} mg$_{\text{prot}}{}^{-1}$]) was the BSH activity measured in the feces collected during the *in vivo* experiments. Additional mechanisms of the $LPBS$ dynamics were the release of hepatic bile salts HBS through the caniculi with rate k_{HBSo} ([day^{-1}])— first step of the enterohepatic circulation. A major part of PBS is reabsorbed in the distal ileum through direct absorption by the epithelium of an emulsion of cholesterol and

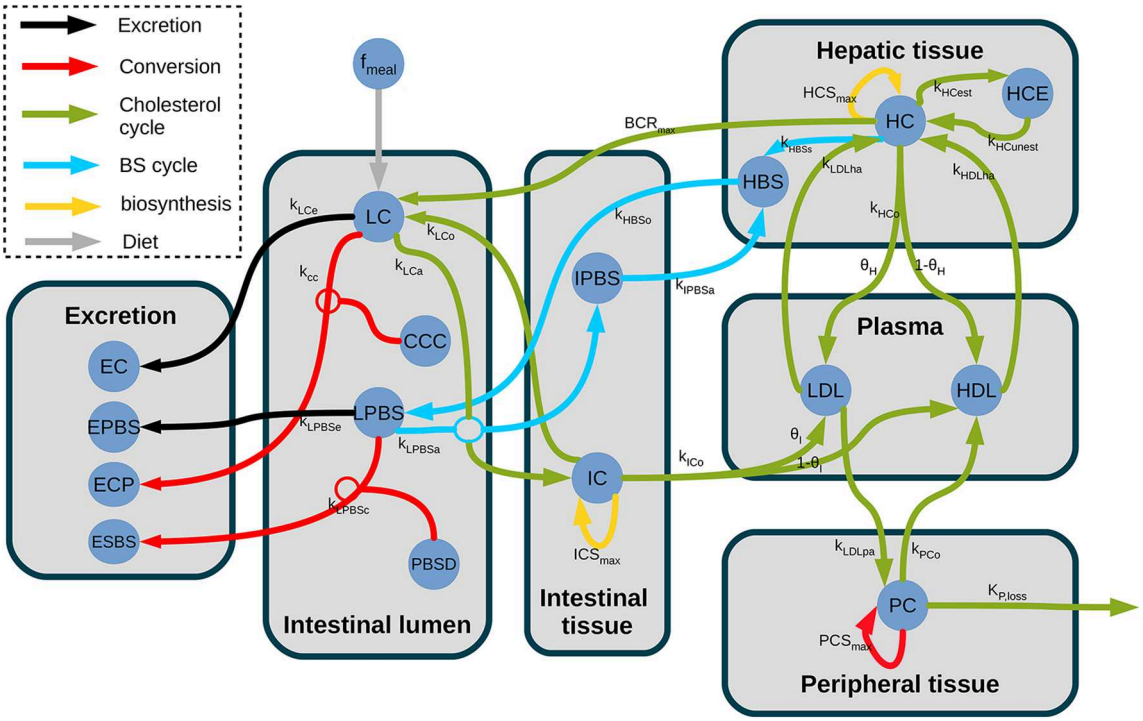


FIGURE 3 | Structure of the model of whole-body cholesterol metabolism. The different compartments included in the model are displayed as gray boxes. The cholesterol flux are indicated by arrows. The gray arrows display the dietary cholesterol influx while the black arrows show the excretion and the orange arrows represent the bacterial transformations. The entero-hepatic BS cycle is displayed in light blue, while the cholesterol cycle is represented in green. The yellow arrows represent the cholesterol biosynthesis. f_{meal}, dietary cholesterol; LC, luminal cholesterol; CCC, coprostanol-to-cholesterol converter; LPBS, luminal primary bile salts; PBSD, primary bile salts converter; EC, excreted cholesterol; EPBS, excreted bile salts; ECP, excreted coprostanol; ESBS, excreted secondary bile salts; IPBS, intestinal primary bile salts; IC, intestinal cholesterol; LDL, low density lipoprotein; HDL, high-density lipoprotein; HC, hepatic cholesterol; HCE, hepatic cholesterol esters; HBS, hepatic bile salts; PC, peripheral cholesterol; k_{LCe}, Luminal cholesterol excretion; k_{cc}, Cholesterol conversion to coprostanol; k_{LPBSe}, Luminal PBS excretion; k_{LPBSc}, Luminal PBS conversion to SBS; k_{LCa}, Luminal cholesterol absorption; k_{LPBSa}, Luminal PBS absorption; k_{LCo}, Epithelial cholesterol secretion in lumen; ICS_{MAX}, Intestinal synthesis maximal rate; k_{ICo}, Intestinal cholesterol outflow; θ_l, Proportion of cholesterol in LDL; k_{IPBSa}, PBS absorption by the liver; k_{HBSo}, BS outflow in lumen; k_{LDLpa}, peripheral absorption in LDL pool; PCS_{MAX}, Peripheral synthesis maximal rate; k_{PCo}, Peripheral cholesterol outflow; $k_{P,loss}$, Cholesterol storage; k_{HCo}, Epithelial cholesterol outflow; θ_H, proportion of cholesterol in LDL; k_{HBSs}, BS synthesis from cholesterol; BCR_{MAX}, Chol. release maximal rate; HCS_{MAX}, Hepatic synthesis max. rate; k_{HCest}, Esterification; $k_{HCunest}$, Unesterification; k_{LDLha}, Hepatic absorption in LDL pool; k_{HDLha}, Hepatic absorption in HDL pool.

BS with rate k_{LCa} ([L mg^{-1} day^{-1}]). A residual excretion through the feces was modeled with the rate k_{LPBSe} ([day^{-1}]). This resulted in the equation

$$\partial_t[LPBS] = \frac{V_H}{V_L}k_{HBSo}[HBS] - k_{LPBSD}[LPBS]PBSD^2 - k_{LCa}[LC][LPBS] - k_{LPBSe}[LPBS]. \quad (8)$$

where V_L and V_H ([L]) were the volumes of the luminal and hepatic compartments.

Luminal cholesterol (LC) dynamics: LC mainly comes from the dietary intake f_{meal} and an hepatic flux through the biliary canal, modulated by the hepatic cholesterol concentration $[HC]$ (Mc Auley et al., 2012). When $[HC]$ is above an hepatic cholesterolemia threshold BCR_t, the flux reaches a maximal rate BCR_{max} while it collapses when the hepatic cholesterol level is below BCR_t. The sensitivity of this regulation is driven by the parameter BS ([−]). An additional influx comes from the intestinal epithelium, with rate k_{LCo}, modulated by $LPBS$ (Van der Velde et al., 2007). Additional sinks are the natural excretion modeled by a constant outflow k_{LCe}, and the cholesterol absorption by the intestinal tissues promoted by the bile salt.

To characterize *in vivo* the cholesterol-to-coprostanol conversion, we used literature data for the ratio $Q_{col,cop} := \frac{[EC]}{[ECP]}$ between excreted cholesterol ($[EC]$) and coprostanol ($[ECP]$) levels in the feces. Low human converters have a ratio $Q_{col,cop} \simeq 0.01$, whereas high human converters have a ratio up to $Q_{col,cop} \simeq 4$ (Sekimoto et al., 1983). We assumed an intermediary conversion ratio by taking $Q_{col,cop} = 0.1$ and we set the conversion time rate $k_{cc} := Q_{col,cop}k_{LCe}$. Furthermore, we properly rescale the K_{D8} Monod constant by taking $K_{CCC} = K_{D8}\frac{[\bar{B}_{spD8}]_{max}}{b_{gut,max}}$. We got at end

$$\partial_t[LC] = \frac{f_{meal}}{V_L} + \frac{V_H}{V_L}\frac{BCR_{max}}{1+\left(\frac{BCR_t}{[HC]}\right)^{BS}} - k_{LCa}[LC][LPBS] + \frac{V_I}{V_L}k_{LCo}[IC][LPBS] - k_{LCe}[LC] - k_{cc}\frac{[LC]CCC}{K_{CCC}+CCC}. \quad (9)$$

3.3.2. Enterohepatic BS Cycle

A part of the *LPBS* is directly excreted into the faecall compartment *EPBS* with rate k_{LPBSe} or is degraded by the BSH producers into the excreted secondary bile salts compartment *ESBS*. The total amount of excreted compounds was followed up, but a density was computed when needed by dividing by the total excretion volume $V_E(t)$ at time t, estimated from the daily stool volume V_{st} with formula $V_E(t) = V_{st}t$. The other part is absorbed together with cholesterol with rate k_{LCa} to constitute an intestinal tissue PBS pool. Then, cholesterol and BS are transported with rate k_{IPBSa} to the liver through the portal vein in order to continue the enterohepatic cycle. In the liver, cholesterol-to-BS biotransformation takes place; it was modeled by an overall transformation rate k_{HBSs} modulated by a negative retro-control of the hepatic bile salts levels *HBS*. We finally got the dynamics of the BS in the excreted compartment *EPBS* and *ESBS*, in the intestinal tissues ([*IPBS*]) and in the liver ([*HBS*]):

$$\partial_t EPBS = V_L k_{LPBSe}[LPBS], \tag{10}$$

$$\partial_t ESBS = V_L k_{LPBSD}[LPBS]PBSD^2, \tag{11}$$

$$\partial_t[IPBS] = \frac{V_L}{V_I}k_{LCa}[LC][LPBS] - k_{IPBSa}[IPBS], \tag{12}$$

$$\partial_t[HBS] = k_{HBSs}\frac{[HC]}{[HBS]} - k_{HBSo}[HBS] + \frac{V_I}{V_H}k_{IPBSa}[IPBS]. \tag{13}$$

We also had $[EPBS](t) = EPBS(t)/V_E(t)$ and $[ESBS](t) = ESBS(t)/V_E(t)$.

3.3.3. Whole-Body Dynamics of Cholesterol

In the lumen, the cholesterol is distributed between the intestinal tissues (through absorption) and the excretion compartment.

3.3.3.1. Excreted cholesterol

A part of the luminal cholesterol is transported into the excreted cholesterol pool (*EC*) in the feces while another part is biotransformed into coprostanol: we assumed that the coprostanol created in the lumen was directly excreted into the feces in the excreted coprostanol pool (*ECP*). Again, we tracked the total amount of excreted components, and recovered density by dividing by $V_E(t)$.

$$\partial_t EC = V_L k_{LCe}[LC], \tag{14}$$

$$\partial_t ECP = V_L k_{cc}\frac{[LC]CCC}{K_{CCC} + CCC}. \tag{15}$$

3.3.3.2. Cholesterol in intestinal tissues

In the intestinal mucosa, additionally to the absorption of the luminal cholesterol and the direct release of cholesterol into the lumen, an endogenous cholesterol synthesis was considered. As in Mc Auley et al. (2012), we assumed that the intestinal tissues activate the cholesterol synthesis when the free cholesterol pool reaches a minimal threshold IC_t. The cholesterol is then produced with a constant rate ICS_{max} and the transition between the production and the resting regimes is modulated by a sensitivity parameter *IS*. Finally, intestinal cholesterol flows toward the plasmatic compartment with a rate k_{ICo}. We got

$$\partial_t[IC] = \frac{V_L}{V_I}k_{LCa}[LC][LPBS] - k_{LCo}[IC][LPBS] + \frac{ICS_{max}}{1 + \left(\frac{[IC]}{IC_t}\right)^{IS}} - k_{ICo}[IC]. \tag{16}$$

3.3.3.3. Plasmatic cholesterol

The cholesterol is transported in the plasma by lipoproteins that are usually separated in distinct lipoproteins populations according to their content of cholesterol and triglycerids. Here, we considered only two lipoproteins compartments which are the most significant for cholesterolemia: high ([*HDL*]) and low density lipoproteins (*LDL*). We considered an absorption flux k_{ICo} (resp. k_{HCo}) from the intestinal (resp. hepatic) tissues which is dispatched into the plasmatic compartments with proportion θ_I (resp. θ_H) for the *LDL* compartment and $(1 - \theta_I)$ (resp. $(1 - \theta_H)$) for the *HDL* compartment. We assumed that the peripheral cholesterol flows in the *HDL* pool only (van de Pas et al., 2011) with rate k_{PCo}. The internal flux between *HDL* and *LDL* pools are reduced to the maturation from high to low density lipoproteins with rate k_{HDLc} (van de Pas et al., 2011). The reverse process occurs with rate k_{LDLc}. We finally modeled outgoing fluxes toward the hepatic and peripheral tissues with, respectively, rates k_{LDLha} and k_{LDLpa} for the *LDL* compartment and k_{HDLha} for the *HDL* carriers (peripheral absorption of *HDL* cholesterol is not included). We then obtained, noting V_B and V_P the volume of the blood and peripheral compartments,

$$\partial_t[HDL] = \frac{V_I}{V_B}(1 - \theta_I)k_{ICo}[IC] + \frac{V_H}{V_B}(1 - \theta_H)k_{HCo}[HC] + \frac{V_P}{V_B}k_{PCo}[PC] - k_{HDLha}[HDL], \tag{17}$$

$$\partial_t[LDL] = \frac{V_I}{V_B}\theta_I k_{ICo}[IC] + \frac{V_H}{V_B}\theta_H k_{HCo}[HC] - (k_{LDLha} + k_{LDLpa})[LDL]. \tag{18}$$

3.3.3.4. Hepatic cholesterol

We separated the liver cholesterol metabolism in three main pathways: (1) an endogenous cholesterol synthesis with parameters HC_t, HCS_{max}, and *HS* like in the intestine; (2), esterification/de-esterification of free cholesterol with conversion rates k_{HCest} and $k_{HCunest}$; (3) ingoing/outgoing flux from the plasma with the rates k_{LDLha}, k_{HDLha}, and k_{HCo}. Hepatic discharge of cholesterol through the canaliculi is modeled with the term $\frac{BCR_{max}}{1+\left(\frac{BCR_t}{[HC]}\right)^{BS}}$ that was introduced in the description of the luminal compartment. This is expressed in Equations (19) and (20)

$$\partial_t[HC] = \frac{V_B}{V_H}k_{LDLha}[LDL] + \frac{V_B}{V_H}k_{HDLha}[HDL] - k_{HCo}[HC] + \frac{HCS_{max}}{1 + \left(\frac{[HC]}{HC_t}\right)^{HS}} - k_{HCest}[HC] + k_{HCunest}[HCE] - k_{HBSs}\frac{[HC]}{[HBS]} - \frac{BCR_{max}}{1 + \left(\frac{BCR_t}{[HC]}\right)^{BS}}, \tag{19}$$

$$\partial_t[HCE] = k_{HCest}[HC] - k_{HCunest}[HCE]. \tag{20}$$

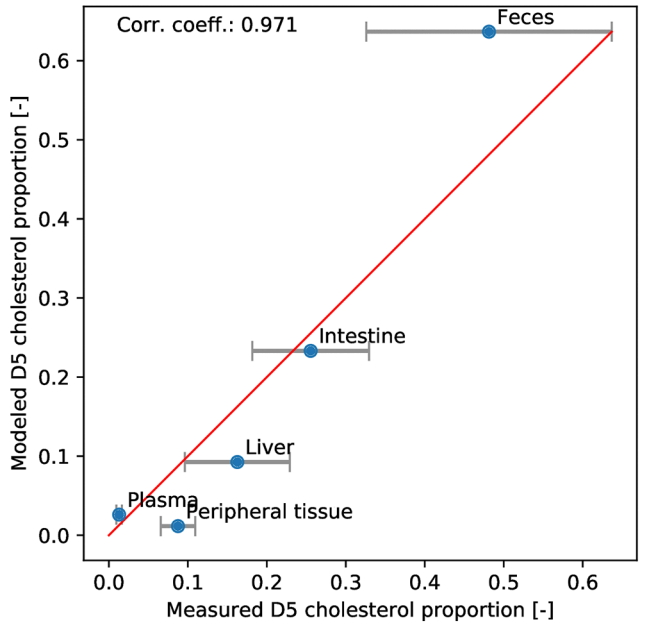


FIGURE 4 | Model validation. The deuterated cholesterol distribution in compartments obtained with the model is plotted against the experimental one. Errorbars representing the SEM of the experimental data are added. We observe that the points follow the $y = x$ line (red) with a high correlation coefficient (0.97).

3.3.3.5. Peripheral cholesterol

Plasmatic cholesterol can be stored in the remaining body tissues, represented by the peripheral cholesterol pool (*PC*). Both *LDL* and *HDL* plasmatic cholesterol are uptaken with rate k_{LDLpa} and k_{LDLha}, respectively. Cholesterol synthesis parameters are PCS_{max}, PC_t, and *PS*. Finally, a global loss is taken into account through the parameter k_{Ploss}, to model storage in adipose tissues. We finally got

$$\partial_t[PC] = \frac{V_B}{V_P}k_{LDLpa}[LDL] - k_{PCo}[PC] + \frac{PCS_{max}}{1+\left(\frac{[PC]}{PC_t}\right)^{PS}} - k_{Ploss}[PC]. \quad (21)$$

All the model parameters (except the bacterial growth model parameters that were inferred as presented in section 3.2), were obtained with steady-state flux and concentration data from the literature (see **Tables S5**, **S6**, and **Table 2**) and the calibration strategy detailed in section 2.7 and in **Table S1**. No additional inference was performed to fit the whole body model with the *in vivo* experimental data.

3.4. Model Validation

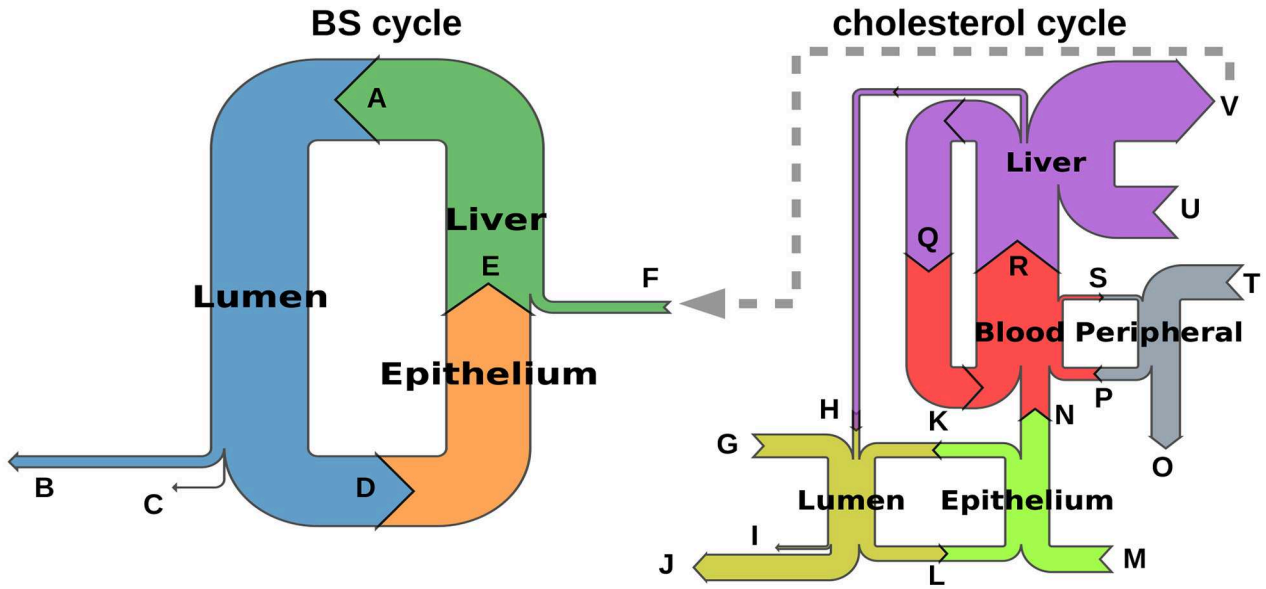
3.4.1. Validation From Deuterated Cholesterol Experimental Data

We used *in vivo* labeled cholesterol data to validate our new model. We duplicated all the cholesterol and BS pools in order to separate the deuterated and normal sterols and monitored their respective dynamics. The resulting model is presented in Equations (S1) to (S28) in the **Supplementary Material**. At initial state, the deuterated components are set to zero in every compartments. Then, the dietary influx of deuterated cholesterol is set to correspond to the experimental levels. After 3 days, the simulation is stopped and the different pools of normal and labeled cholesterol and BS are recomposed to reconstruct the intestinal, excreted, plasmatic, peripheral and hepatic levels of normal and labeled cholesterol. Then, the distribution obtained with the model is compared to the experimental distribution (**Figure 4**). We observed that the points of the scatter plot followed the $y = x$ line with a correlation coefficient of 0.97. This strong agreement between model and data indicated that the model correctly captured the flux between the different compartments.

3.4.2. Flux Repartition at Steady State

We computed a basal simulation until steady state and observed the resulting flux between compartments. As expected, we recovered the steady state flux from published data that were used for the model calibration (see **Tables S5**, **S6**, and **Table 2**, Supplementary Data). We represented the flux in a Sankey graph (**Figure 5**) of the cholesterol and BS whole body cycles. The Sankey graph helped visualizing mass transfers since it displayed the flux distribution with arrows proportional to the flux that they represent. The large discrepancy between BS and cholesterol flux was particularly emphasized with this representation. For example, while the BS biosynthesis (ss_{kHBS_s} in the model) is a major sink for the cholesterol cycle, it only represents a minor influx for the BS cycle, counterbalancing the small BS excretion (**Figure 5**, gray dashed arrow). The BS pool conservation mainly relies on BS recycling, which is fueled by large absorption and transport capacities in the lumen, the intestine and the liver. The basal bacterial conversion to SBS represents a negligible outflux compare with the BS circulation (**Figure 5**, left).

We observed that the cholesterol cycle was roughly separated in three main entities (**Figure 5**, right). (i) A central axis (intestinal epithelium-blood system-liver axis) supports the main part of cholesterol transfers. (ii) The luminal compartment represents the second cholesterol route; it is connected to the central axis by the epithelial interface and the biliary cholesterol release. The net balance of the cholesterol exchanges with the main central stream is slightly negative: the cholesterol absorption by the epithelium is counterbalanced by the cholesterol secretion while the small biliary cholesterol release supports the luminal cholesterol levels. Again, the basal cholesterol-to-coprostanol bacterial conversion is secondary. (iii) The third entity is composed by the peripheral tissues. In this compartment, the cholesterol biosynthesis is nearly entirely balanced by the cholesterol storage in adipose tissues, giving a slightly positive contribution to the main central cholesterol flux. In the central axis, the BS biosynthesis is by far the principal outflux of the cholesterol cycle, and is mainly fueled by the hepatic and epithelial cholesterol biosynthesis. The two-side cholesterol exchanges between the liver and the blood constitute an important cholesterol sub-cycle: this loop could be seen as a buffer that regulate the BS biosynthesis outflux, by absorbing cholesterol fluctuations.



Ref.	Flux	value [mg day^{-1}]	Description
A	SS_{kHBSo}	20.00	BS release
B	SS_{kLPBSe}	2.71	BS excretion
C	SS_{kLPBSc}	0.00	Bact. conversion
D	SS_{kLPBSa}	17.28	Epith. absorption
E	SS_{kIPBSa}	17.28	Transport
F	SS_{kHBSs}	2.72	BS synthesis

Ref.	Flux	value [mg day^{-1}]	Description
G	$SS_{k_{in}}$	0.78	Dietary intake
H	$SS_{BCR_{max}}$	0.19	Chol.release
I	SS_{kcc}	0.09	Bact. conversion
J	SS_{kLce}	0.87	Excretion
K	SS_{kLCo}	0.49	Secretion
L	SS_{kLCa}	0.49	Absorption
M	$SS_{ICS_{max}}$	0.87	Synthesis
N	SS_{kICa}	0.88	Transport
O	SS_{kPloss}	0.88	Stock.
P	SS_{kPCo}	0.42	Outflow
Q	SS_{kHCo}	1.38	Outflow
R	SS_{B_H}	2.55	Transport
S	SS_{kLDLpa}	0.14	Absorption
T	$SS_{PCS_{max}}$	1.16	Synthesis
U	$SS_{HCS_{max}}$	1.75	Synthesis
V	SS_{kHBSs}	2.72	BS synsthesis

FIGURE 5 | Sankey diagrams of the BS and cholesterol cycles. We display the Sankey diagrams of the BS and cholesterol cycles at steady state. Each row is proportional to the corresponding flux (mg day^{-1}), and is displayed with a letter referring to the corresponding model coefficient, its steady-state value and its nomenclature in the model, gathered in the tables. We note that there is a huge discrepancy of flow magnitude between the two cycles, the BS cycle involving much more higher mass transfers than the cholesterol one. Thus, we could not represent the diagrams with the same scale, resulting in different arrow thicknesses for the BS synthesis, despite an equal value for this flux in the two cycles. We emphasize this scale change and the connection between both cycles with the gray dashed arrow. Flux details can be found in **Table S6** and **Table 2**.

3.5. Numerical Exploration of the Bacterial Impact on Cholesterolemia

To illustrate the impact of bacterial metabolism on the whole-body cholesterol cycle and to provide a first analysis of the mechanisms involved, we performed three new simulations enhancing, respectively, (i) the bacterial carrying capacity of the BS converters, (ii) the cholesterol converters, or (iii) both. Namely, we multiplied by 20 the $PBSD_{MAX}$ (resp. CCC_{MAX}) parameter which represents a 20-fold growth of the corresponding population, i.e., a small bacterial increase compared to the several log fold changes that can occur during bacterial colonization of the digestive track. We then displayed the corresponding Sankey graphs of the steady state BS and cholesterol cycles (**Figure S3**) with bar plots (**Figure 6**, **Supplementary Material**) representing the relative variations comparatively to the basal simulation of the different flux and pool concentrations.

The enhancement of the BS converter populations *PBSD* increased the bacterial activity which dropped down the luminal level of BS by about 26% (**Figure 6**, bottom, LPBS). This reduction induced a 28% decrease of the epithelial absorption of luminal BS, but also of luminal cholesterol (**Figure 6**, top, ss_{kLPBS_a}). In the mean time, the cholesterol intestinal excretion was decreased so that the net balance of cholesterol exchanges with the central axis was only slightly reduced (**Figure S3**, top), buffering the reduction of the cholesterol absorption and reducing its impact on the whole-body cholesterol cycle. However, the decrease of the BS epithelial absorption had stronger effects on the cholesterol regulation. To counterbalance this loss, the BS biosynthesis was increased by about 17% (**Figure 6**, top, ss_{kHBS_s}), fueled by a 23% growth of the liver cholesterol biosynthesis. Worthy of note, the contribution to the cholesterol cycle of the intestinal biosynthesis remained unchanged, whereas the liver-plasma exchanges were reduced by 9% to free up cholesterol for the BS biosynthesis. HDL and LDL cholesterol concentrations decreased by about 5%.

The increase of the CCC population had a different impact on the cholesterol and BS cycles. The higher loss of cholesterol in the lumen by direct excretion or conversion into coprostanol led to a huge decrease (47%) of the luminal cholesterol level (**Figure 6**, bottom, LC) which reduced by 35% the cholesterol and BS absorption, leading to a 21% increase of BS level in the lumen (**Figure 6**, top, ss_{kLPBSa}, and bottom, LPBS). In turn, higher luminal BS level increased the excretion and promoted the intestinal cholesterol secretion, inducing a net negative cholesterol flux from the intestinal epithelium to the lumen (**Figure S3**, bottom right). This local reduction of cholesterol influx in the intestinal epithelium was partially balanced by a stronger intestinal cholesterol synthesis by 19%, but the net contribution of the intestinal tissues to the central cholesterol stream was reduced by 0.06 mg day^{-1} comparatively to the basal

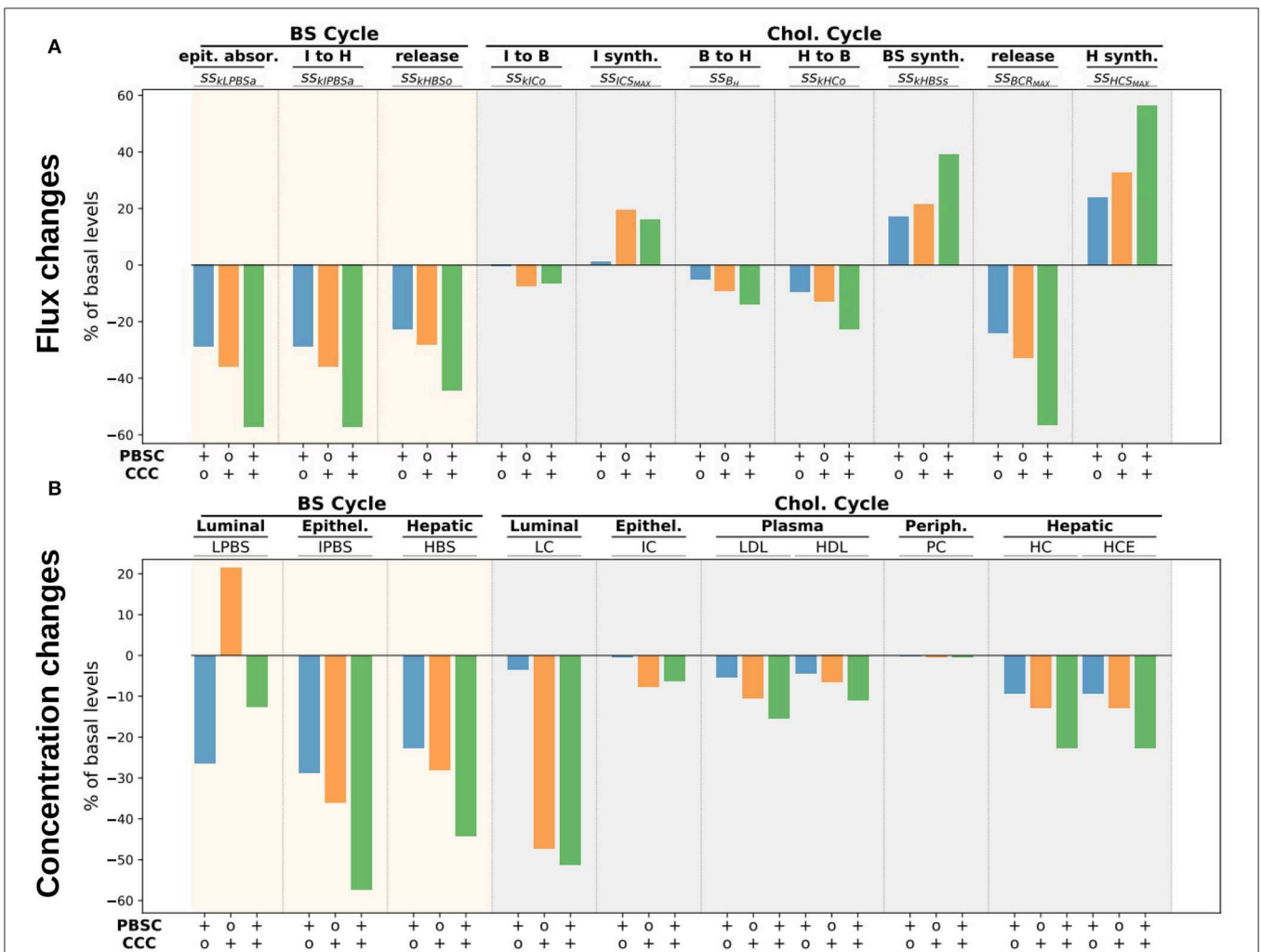


FIGURE 6 | Flux and concentration changes for higher bacterial activity. We display flux **(A)** and concentration **(B)** changes (in percentage of the basal respective quantities) for a 20-fold increase of PBSD (resp. CCC) levels in the lumen, i.e., BS (resp. cholesterol) bacterial converters. The steady state flux nomenclature can be found in **Table S6** and **Table 2**.

activity (**Figure S3**, bottom right, and **Figure 5**), showing that the conversion of luminal cholesterol had a direct impact on the cholesterol cycle. In addition to this direct action on the central cholesterol stream, the same indirect mechanism that took place in the high *PBSD* experiment occurred. The reduced BS absorption was compensated by a higher BS biosynthesis, with a higher magnitude (21% increase for high *CCC* vs. 17% increase for high *PBSD* populations). Again, the BS biosynthesis increase was allowed by a higher cholesterol hepatic biosynthesis (32%) and by a reduced transport between the liver and the blood (12%, **Figure 6**, top, $ss_{HCS_{MAX}}$ and ss_{kHCo}). We observed that the magnitude of the flux involved in the indirect BS-mediated regulation of the cholesterol was higher than the direct loss of cholesterol allowed by the bio-conversion of cholesterol to coprostanol (**Figure S3**, bottom right). The impact on the plasmatic cholesterol levels was also more important, with a 6.6 and 10.4% reduction for the HDL and LDL, respectively.

When the two bacterial functions were both enforced, the mechanisms tended to sum up, leading to a lower BS and cholesterol absorption by the epithelium (approximatively a 60% decrease) and an increase of the BS synthesis by 40%. The plasmatic levels of HDL and LDL reduced by 10.9 and 15.3%, respectively. We finally observed that the impact on the peripheral cholesterol was very weak in the three cases.

3.6. Local and Global Sensitivity Analysis of the Model

After this first exploration of the bacterial impact on the cholesterol fate, we went deeper in the analysis by conducting a systematic numerical exploration. We first conducted a local sensitivity analysis of the model, relatively to the bacterial converter carrying capacities in the gut microbiota, in order to study the model response when the bacterial levels evolved. Then, we performed a global sensitivity analysis by shifting the parameters that govern eleven flux of the BS and cholesterol cycles, in order to study the relative importance of each flux in the output variability.

3.6.1. Local Sensitivity Analysis

We present the result of the local sensitivity analysis in **Figure S4**, where different fluxes and concentrations variations were plotted against log-fold changes of the bacterial carrying capacities of the CCC (orange lines, crosses) and PBSD (blue lines, circles) populations, comparatively to the basal carrying capacities. We observed that decreasing the bacterial levels had little impact on the overall behavior of the model. When the cholesterol converters carrying capacity was weaker, a slight increase of cholesterol levels (luminal cholesterol LC, intestinal cholesterol IC, LDL, **Figure S4**) was observed, but smaller BS converter levels had no effect on the cholesterol or BS cycles due to the negligible basal BS conversion (**Figure 5**). Conversely, a monotonous evolution of the different flux and concentrations was observed when the bacterial populations levels were increased. No saturation effects could be observed.

Several features previously observed in **Figure S3** and **Figure 6** for a 20-fold increase were confirmed. When the cholesterol conversion activity was enhanced, we observed a constant increase of luminal BS concentration, together with a decrease of the BS intestinal absorption (*LPBS* and ss_{LPBSa}, **Figure S4**). Varying PBSD levels had a very limited impact on the intestinal cholesterol, on the transport from the intestinal tissues to the blood stream and on the intestinal cholesterol synthesis (*LC*, *IC*, ss_{kICo}, and $ss_{ICS_{MAX}}$, **Figure S4**). This observation enforced the claim that the interaction of the BS conversion with the cholesterol cycle mainly occurred through the BS synthesis, and not through a direct variation of the cholesterol absorption. Finally, the impact of the bacterial activity on peripheral cholesterol remained very weak whatever bacterial level (*PC*, **Figure S4**).

The bacterial effect on the whole-body cholesterol and BS cycles varied differently when the *CCC* and *PBSD* carrying capacities changed. For intermediate bacterial concentrations (1 log-fold change comparatively to the basal levels), the cholesterol converters provided higher effects on the cholesterol and BS pools. But for higher bacterial levels (2 log-fold change), the BS converters had a stronger impact on the different flux and concentrations that were observed, except in the intestinal tissue compartment where the luminal *BS* modulation induced by the bacterial converters had little effects (*LC*, *IC*, ss_{kICo}, and $ss_{ICS_{MAX}}$, **Figure S4**). We noted that the variations reached 50% for the highest CCC population in the luminal cholesterol compartment, while for the highest BS converter population, this level of variation is obtained all along the enterohepatic cycle (*LPBS*, *IPBS*, *HBS*) and for the hepatic cholesterol concentrations (*HC* and *HCE*).

3.6.2. Global Sensitivity Analysis

A global sensitivity analysis was performed by modifying 11 parameters controlling the flux involved in the enterohepatic BS cycle and the whole-body cycle of cholesterol including the dietary cholesterol intake (k_{in}), the biliary cholesterol release (BCR_{MAX}), the luminal cholesterol absorption (k_{LCa}), the cholesterol transport from the liver to the blood ($kHCo$) and the reverse flux ($B \rightarrow H$, sum of k_{LDLha} and k_{HDLha} that were shifted simultaneously), the cholesterol synthesis (by shifting at the same time the ICS_{MAX},HCS_{MAX}, and PCS_{MAX} parameters driving, respectively, the intestinal, hepatic and peripheral cholesterol biosynthesis), the cholesterol and BS epithelial absorption (k_{LPBSa}), the BS release in the lumen (k_{HBSo}), the BS biosynthesis (k_{HBSs}) and the bacterial population carrying capacities ($PBSD_{MAX}$ and CCC_{MAX}). We displayed the Sobol first order index, and the PCC of the different parameters for the concentration outputs in each compartment (namely, the luminal LPBS, the intestinal epithelium IPBS, the hepatic HBS levels for the BS cycle, and the luminal LC, epithelial IC, plamatic HDL and LDL, peripheral PC and hepatic HC and HCE cholesterol pools). The Sobol index measures the contribution of a given parameter to the variability of the observed output while PCC quantifies the correlation between parameter and output variations. Both criteria are complementary: while the former helps identifying the main drivers of a given output the later also provides feedback on the sign of the interaction between parameter and output. The total sum of the Sobol indices was nearly 1 for almost all compartments, indicating that the total

variance was entirely explained by the individual variation of the parameters tested. However, for the compartments modeling the enterohepatic cycle and the LC pool a residual variance was observed, meaning that parameter interactions contributed significantly to the total variance.

As expected, the bacterial carrying capacities had a stronger negative impact on the concentration of their respective substrates in the lumen, i.e., *LPBS* (resp. *LC*) for the *PBSD* population (resp. *CCC*). The dietary intake also positively impacted the luminal cholesterol *LC* but had very little influence on the other compartments. We noted that the *PBSD* population was the main parameter that tuned down the whole enterohepatic cycle, whereas the effect of the *CCC* population was concentrated on the luminal compartment, the main (positive) contributor to the cholesterol cycle variations being the cholesterol biosynthesis. A notable impact of the BS deconjugation on the hepatic cholesterol concentrations was detected. It must be related to the strong variations noticed for *LPBS* in the local sensitivity analysis (**Figure S4**). Interestingly, the impact on hepatic cholesterol variations was distributed among several parameters, mainly biosynthesis, BS production, BS release and *PBSD* populations activity, all being negative but the cholesterol biosynthesis.

The main driver of the LDL and HDL plasmatic levels, which are the main biomarkers for cholesterolemia, was the hepatic cholesterol absorption: the most efficient way to reduce plasmatic cholesterol was enhancing the transport between the plasma and the liver. The cholesterol biosynthesis by the different organs and the transport from the liver to the plasma came in second and third position. The cumulative bacterial contribution was small and occupied the fourth rank, with an impact similar to the BS biosynthesis or the BS release. While the impact of the cholesterol converters was minor, the BS converters supported the main part of the bacterial contribution to the plasmatic cholesterol levels. This impact was up to a 27 and 49% reduction for, respectively, the LDL and hepatic cholesterol for a 2-log increase of BS converter levels (see **Figure S4**).

4. DISCUSSION

4.1. Mathematical Modeling Provided Improved Insights in the Cholesterol Cycle

In system biology, mathematical models can be used to link heterogeneous data taken at different scales. Modeling allows to connect these observations with a sequence of mechanisms involved in regulatory processes, enabling the co-interpretation of the data otherwise difficult to achieve without the model.

Here, we used a mathematical model to interpret together *in vitro* bacterial activity with *in vivo* animal experiment data. The *in vitro* model provided a quantitative evaluation of bacterial uptake and production rates on BS and cholesterol, which was upscaled to represent the microbial activity in the small intestine. This microbial metabolism was then plugged into a whole-body model of cholesterol and BS cycle to study the systemic impact of the different cycle drivers. This whole-body model was derived from existing models. The overall structure and rate expression of the main mechanisms was taken from Mc Auley et al. (2012) and Morgan et al. (2016), and substantially simplified according to van de Pas et al. (2010). Compared to Morgan et al. (2016), the very detailed description of cholesterol metabolism was simplified by keeping primary and final metabolites only. An accurate population model of cholesterol transport in lipoprotein has been developed in Sips et al. (2014), that we summed up by considering only two lipoprotein compartments: HDL and LDL. The model was calibrated using the method and values taken from van de Pas et al. (2011). An additional cholesterol outflow has been added from the intestinal tissue into the lumen, as observed and measured in Van der Velde et al. (2007). The outputs of the complete model were compared to the animal experiment data. As a whole, this modeling approach allowed to integrate the different data in a comprehensive framework and showed the consistency of the modeled mechanisms with the experiments.

The model provided a simplified description of cholesterol distribution at steady-state. The BS cycle appeared to be well-balanced, showing similar flux levels across its different components in a Sankey graph (see **Figure 5**). Unlike BS cycle, the cholesterol cycle presents an uneven repartition, the flux crossing the liver and the blood being sensibly higher than those involved in the other compartments (see **Figure 5**). This systemic view suggests that BS biosynthesis is the principal cholesterol flux, mainly supported by cholesterol synthesis in the liver, and by a buffering pool composed by cholesterol exchanges between the blood and the liver. This simplified view allows one to hypothesize that blood cholesterol levels will be mainly driven by the transport mechanisms between the blood and the liver, whereas liver cholesterol reduction could be strongly impacted by the biosynthesis of cholesterol (positively) and BS (negatively). In the cholesterol and BS cycles, the bacterial fluxes are small compared to others. But as BS fluxes are one order of magnitude higher than cholesterol fluxes, a small sink flux in the BS cycle can have a significant impact in the cholesterol cycle, making bacterial BSH activity a potential effective driver of cholesterol levels in the hosts.

In depth numerical exploration of the model allowed ranking the main factors that influence the distribution of cholesterol in the body. Global sensitivity analysis confirmed the actual effect of bacterial activity on host cholesterolemia (see **Figure 7**). If the impact of cholesterol-to-coprostanol conversion on the overall cholesterol cycle was small, bacterial BS conversion had greater effect on the liver cholesterol level. Plasmatic levels proved to be massively controlled by host mechanisms (mainly transport between blood and liver compartments closely followed by cholesterol biosynthesis), whereas bacterial activity impacts as strongly as other host mechanism the hepatic cholesterol pool. We note that the importance of cholesterol transport for plasmatic cholesterol regulation has already been highlighted by both modeling and experimental studies (Field and Gibbons, 2000; Morgan et al., 2016). The model then helped to predict the effect of targeting specific mechanisms to manage the different cholesterol pools, and to sort them by efficiency.

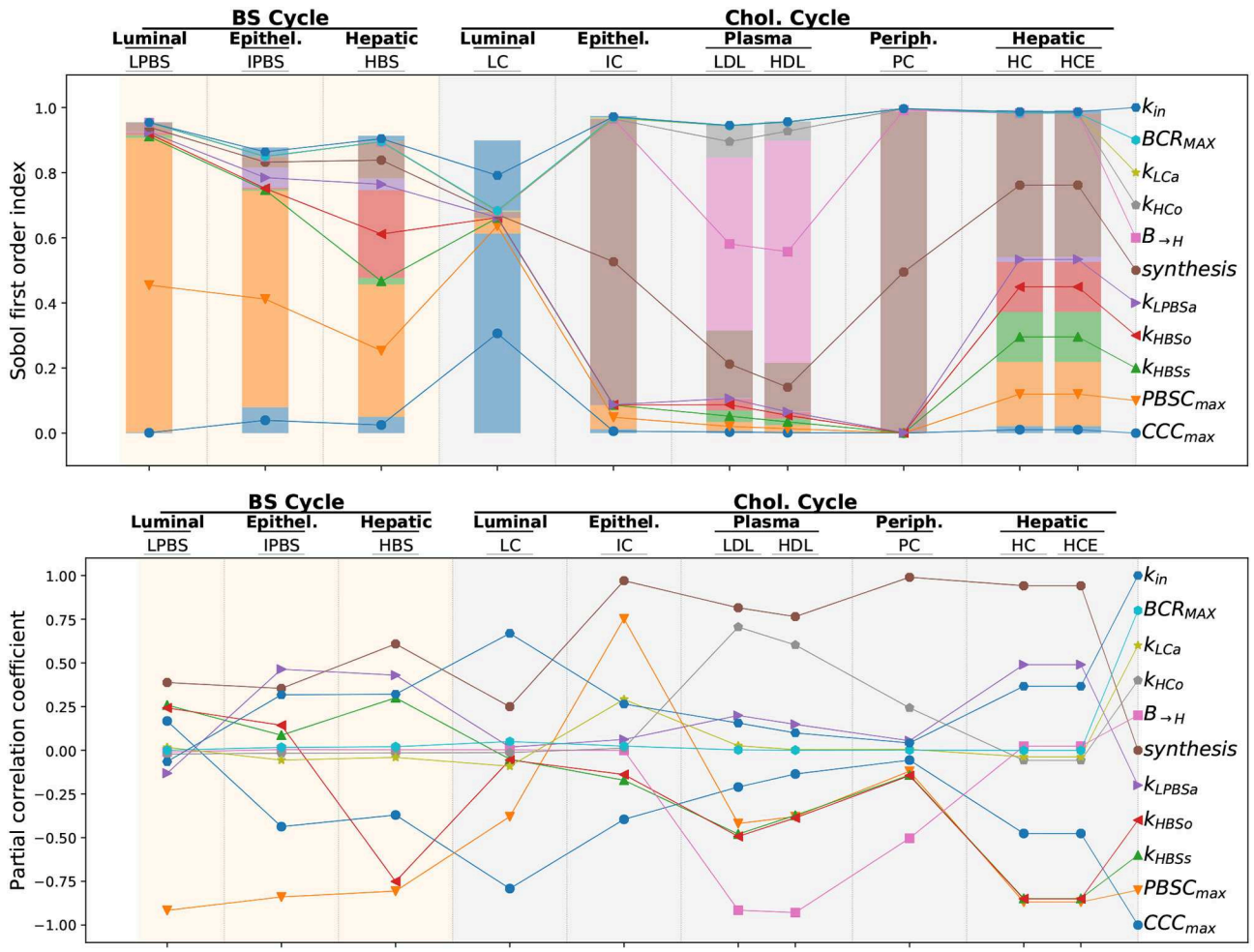


FIGURE 7 | Global sensitivity analysis of steady-state levels of cholesterol and BS in the different compartments. We display, for each steady-state level of cholesterol or BS in the different compartments, the first order Sobol index (top) and the Partial Correlation Coefficient (PCC, down) of the different flux parameters involved in the global sensitivity analysis. The Sobol index measures the proportion of the output variance generated by the variations of a given parameter while the PCC quantifies the correlation between parameter and output variations. In the upper plot, the lines only link together the bar fractions corresponding to the same parameter, in order to facilitate the reading of the figure. The nomenclature is: k_{in}, dietary cholesterol intake; BCR_{MAX}, biliary cholesterol release; k_{LCa}, epithelial cholesterol absorption; *kHCo*, cholesterol transport from the liver to the blood; $B \rightarrow H$, cholesterol absorption by the liver and the reverse flux; *synthesis*, cholesterol synthesis driven by the ICS_{MAX},HCS_{MAX} and PCS_{MAX} parameters; k_{LPBSa}, BS epithelial absorption; k_{HBSo}, BS release; k_{HBSs}, BS biosynthesis; $PBSD_{MAX}$, BS bacterial converters; CCC_{MAX}, cholesterol bacterial converters.

4.2. Limitations and Potential Improvements for Model Validation

Some assumptions have been made during the model construction that are important to keep in mind for correct interpretation. A first limitation is that the model has been built on mice data: all the flux and steady-state values used for model calibration (see **Table S1**) have been picked up in mice studies, as well as the model validation data taken from our animal model. The insights in regulation mechanisms obtained during this study are valid for mice, and the transposition to humans would need further studies.

Our model entails a drastic reduction of microbiota and host physiology complexity. In this study, the individual activity of two selected bacterial strains with known cholesterol or BS activity was assumed to be representative of the overall activity of a complex microbiota after rescaling, and included in the whole body model. A more realistic mechanistic model of the bacterial activity related to cholesterol metabolism in a complex microbiota would ideally require an ecological model able to track the bacterial phenotypic diversity and interactions with the environment through metabolic models including the relevant metabolic pathways, as this was done for fiber degradation (Muñoz-Tamayo et al., 2010; Labarthe et al., 2019). The complexification and validation of the microbiota model would necessitate the dynamic screening *in vivo* of the BSH activity and cholesterol-to-coprostanol conversion of a complex microbiota. This could be achieved through multi-Omics analyses of feces.

Metagenomic data would indicate the metabolic potential of the microbial community regarding cholesterol metabolism, metatranscriptomic data would give the metabolic activity effectively expressed and targeted metabolomics would show the dynamics of key metabolites (e.g., BS, cholesterol, coprostanol). Our analysis suggests that BS metabolism could be the main target.

In the same way, host physiology has been sketched: we chose to provide as well simple phenomenological models of cholesterol host metabolism. Whereas complete metabolic pathways include a cascade of elementary reactions, we only modeled the global resulting relationship between raw substrate and final metabolites. Here again, odel validation could be completed with additional experiments. As our model is not static, model calibration and validation require both steady state pool values, to capture the physiological levels of the different pools, and flux values between compartments, to describe the regulation processes. Measuring fluxes experimentally is challenging since it necessitates several time points with dedicated reporters, inducing multiple animal sacrifices and significant replicates to mitigate inter-individual variability. That is why we chose to rely on published data for model calibration (for both flux and steady-states), and to check the consistency between the model predictions and the observed distribution of ingested cholesterol after 3 days. Actually, screening labeled cholesterol fate in the host tissue provides a much better picture of the system dynamics than measuring steady state levels only. Indeed, steady state levels could possibly be reproduced by the model if the compartment net fluxes were null, even with inaccurate fluxes between compartments. On the contrary, a correct distribution of labeled cholesterol after 3 days requires correct fluxes, otherwise D5-cholesterol propagation between compartments would not be correctly modeled. The animal experiments then allowed to both validate fluxes and steady-state values, and represented a good balance between experimental load and significance for model validation.

4.3. Is Bacterial Activity an Effective Driver of Cholesterolemia Control?

Functional characterization of bacteria isolated from gut microbiota samples allowed to identify functions related to cholesterol and BS turn-over. The main microbial mechanisms for cholesterol loss that were identified are direct cholesterol biotransformation into coprastanol, BS deconjugation and cholesterol incorporation into microbial membranes (Kriaa et al., 2019), which make the microbial communities a potential driver of cholesterol regulation. However, a classical counter-argument being raised is the spatial segregation between cholesterol and BS absorption, mainly located in the small intestine, and the bacterial populations, mainly located downstream in the large intestine: microbial communities could hardly be an important actor of cholesterol management if they do not have a physical access to cholesterol and BS substrates in order to degrade it before absorption by the human host.

We addressed this issue in two ways. First, we experimentally checked that cholesterol and BS were available in the large intestine by measuring in mice labeled sterol levels in the caecum and the large intestine 3 days after ingestion of the labeled cholesterol. Caecal and colonic cholesterol represented 4.5% of the overall labeled cholesterol. It demonstrates that cholesterol is available to colon microbiome and is present in luminal content and intestinal tissues. Second, we calibrated the bacterial activity of BS deconjugation to be representative of microbial populations located in the small intestine, smaller than colic populations but active. Indeed, we selected the scaling parameter $b_{gut,max}$ of BS deconjugation activity which represents the nominal bacterial concentration, as a proxi of the bacterial levels measured in the small intestine. Furthermore, BSH production is involved in BS tolerance by bacteria (Begley et al., 2005) and may be active in the upper part of the intestinal track where BS levels are high. This was taken into account in the model by mimicking the activity and functional dynamics of the *Bacteroides xylanisolvens* XB1A strain, a BS deconjugation specialist.

4.4. Relative Impact of Host and Bacterial Pathways in Cholesterol Metabolism

The contribution of bacterial pathways to the global cholesterol and BS regulation is complex. Bacterial metabolism is the main driver impacting BS turn-over. On the contrary, the impact of the bacteria on the epithelial and peripheral cholesterol is relatively weak compared to cholesterol biosynthesis by the host. To manage plasmatic and hepatic cholesterol pools, more drivers are available. If transport between blood and liver compartment is the preponderant factor of plasmatic cholesterol variations, the contribution of bacterial pathways is not null. In the liver, the impact of the bacterial pathways have the same order of magnitude than other flux, such as BS production, BS release or cholesterol biosynthesis. Hence, managing the host microbiota to enhance BS and cholesterol conversions in the lumen qualifies as a promising tool to control hepatic, and to a lower extent plasmatic cholesterol, in addition to the usual strategies aiming at controlling cholesterol synthesis and transport between compartments.

5. CONCLUSION

We derived a whole body model of cholesterol dynamics that includes microbial metabolism. This model, based on existing models lacking bacterial compartment, is grounded by *in vitro* experiments to capture the bacterial conversion of BS and cholesterol, and by *in vivo* experiments with labeled cholesterol that allowed model validation. The labeled cholesterol provided a snapshot of the deuterated cholesterol distribution after 3 days, and the model gave a precise view of the flux between compartments in the whole cholesterol and BS cycles. This study showed that cholesterol conversion to BS is the main flux of cholesterol cycle, making bacterial BS degradation a promising target for cholesterol management. An extensive model exploration confirmed numerically the impact of the bacterial activity, and the greater influence of BS degradation

on plasmatic cholesterol levels for high converters. Finally, a global sensitivity analysis indicated that transport from plasma to liver is the main driver of plasmatic cholesterol reduction, but that BS degradation is in second position, with the other BS cycle drivers: BS biosynthesis and BS release in the lumen. Bacterial activity is then a promising additional therapeutic strategy able to provide alternatives for non-responders to existing therapies.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The animal study was reviewed and approved by Animal Care Committee (C2E-45 COMETHEA) INRA, Centre de Jouy-en-Josas, France.

AUTHOR CONTRIBUTIONS

MB and MR performed the *in vitro* and *in vivo* experiments. MB and AK produced the experimental data. ML and PL performed the mass spectrometry analysis. MB and SL developed the mathematical model. SL performed the numerical explorations and produced the figures. MB, SL, and MR drafted the paper. SL, PG, BL, EM, and MR contributed to the conception, design, and funding of the study. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

Funding for this work was from INRA metaprogramme MEM (Modchocycle project), SATT Paris-Saclay Micro-chol project 2016-1211l. AK was supported by the project CMCU-PHC Utique (no 19G0819)-Campus France (no 41786NC). This publication has been written with the support of the AgreenSkills+ fellowship programme which has received funding from the EU's Seventh Framework Programme under grant agreement N° FP7-609398 (AgreenSkills+ contract).

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Dr. Aline Potiron for her helpful discussion regarding the analysis of coprostanol producing bacteria. This work was supported by the Microbiology and the Food Chain division (MICA) of the INRAE Institute through the Metaprogramme MEM - Meta-omics and microbial ecosystems.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01121/full#supplementary-material

REFERENCES

Arnold, D., and Kwiterovich, P. (2003). "Cholesterol-absorption, function and metabolism," in *Encyclopedia of Food Sciences and Nutrition, Vol. 2*, eds B. Caballero, C. L. Trugo, and M. P. Finglas (Amsterdam: Academic Press), 1226–1237. doi: 10.1016/B0-12-227055-X/00225-X

Bazett, M., Bergeron, M.-E., and Haston, C. K. (2016). Streptomycin treatment alters the intestinal microbiome, pulmonary T cell profile and airway hyperresponsiveness in a cystic fibrosis mouse model. *Sci. Rep*. 6:19189. doi: 10.1038/srep19189

Begley, M., Gahan, C. G., and Hill, C. (2005). The interaction between bacteria and bile. *FEMS Microbiol. Rev*. 29, 625–651. doi: 10.1016/j.femsre.2004.09.003

Begley, M., Hill, C., and Gahan, C. G. (2006). Bile salt hydrolase activity in probiotics. *Appl. Environ. Microbiol*. 72, 1729–1738. doi: 10.1128/AEM.72.3.1729-1738.2006

Doré, J., Multon, M.-C., Behier, J.-M., Affagard, H., Andremont, A., Barthelemy, P., et al. (2017). The human gut microbiome as source of innovation for health: which physiological and therapeutic outcomes could we expect? *Therapie* 72, 21–38. doi: 10.1016/j.therap.2016.12.007

Field, P. A., and Gibbons, G. F. (2000). Decreased hepatic expression of the low-density lipoprotein (LDL) receptor and LDL receptor-related protein in aging rats is associated with delayed clearance of chylomicrons from the circulation. *Metabolism* 49, 492–498. doi: 10.1016/S0026-0495(00)80014-1

Folch, J., Lees, M., and Sloane Stanley, G. (1957). A simple method for the isolation and purification of total lipides from animal tissues. *J. Biol. Chem*. 226, 497–509.

Gérard, P., Lepercq, P., Leclerc, M., Gavini, F., Raibaud, P., and Juste, C. (2007). *Bacteroides* sp. strain d8, the first cholesterol-reducing bacterium isolated from human feces. *Appl. Environ. Microbiol*. 73, 5742–5749. doi: 10.1128/AEM.02806-06

Gylling, H. (2004). Cholesterol metabolism and its implications for therapeutic interventions in patients with hypercholesterolaemia. *Int. J. Clin. Pract*. 58, 859–866. doi: 10.1111/j.1742-1241.2004.00351.x

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). Dram: Efficient adaptive MCMC. *Stat. Comput*. 16, 339–354. doi: 10.1007/s11222-006-9438-0

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli* 7, 223–242. doi: 10.2307/3318737

Igel, M., Giesa, U., Lütjohann, D., and von Bergmann, K. (2003). Comparison of the intestinal uptake of cholesterol, plant sterols, and stanols in mice. *J. Lipid Res*. 44, 533–538. doi: 10.1194/jlr.M200393-JLR200

Ikonen, E. (2008). Cellular cholesterol trafficking and compartmentalization. *Nat. Rev. Mol. Cell Biol*. 9:125. doi: 10.1038/nrm2336

Iooss, B., and Lemaître, P. (2015). "A review on global sensitivity analysis methods," in *Uncertainty Management in Simulation-Optimization of Complex Systems*, eds G. Dellino and C. Meloni (Boston, MA: Springer), 101–122. doi: 10.1007/978-1-4899-7547-8_5

Iqbal, J., and Hussain, M. M. (2009). Intestinal lipid absorption. *Am. J. Physiol. Endocrinol. Metab*. 296, E1183–E1194. doi: 10.1152/ajpendo.90899.2008

Jakulj, L., van Dijk, T. H., de Boer, J. F., Kootte, R. S., Schonewille, M., Paalvast, Y., et al. (2016). Transintestinal cholesterol transport is active in mice and humans and controls Ezetimibe-induced fecal neutral sterol excretion. *Cell Metab*. 24, 783–794. doi: 10.1016/j.cmet.2016.10.001

Jones, M., Martoni, C., and Prakash, S. (2012). Cholesterol lowering and inhibition of sterol absorption by *Lactobacillus reuteri* ncimb 30242: a randomized controlled trial. *Eur. J. Clin. Nutr*. 66:1234. doi: 10.1038/ejcn.2012.126

Joyce, S. A., Shanahan, F., Hill, C., and Gahan, C. G. (2014). Bacterial bile salt hydrolase in host metabolism: potential for influencing gastrointestinal microbe-host crosstalk. *Gut Microbes* 5, 669–674. doi: 10.4161/19490976.2014.969986

Kriaa, A., Bourgin, M., Potiron, A., Mkaouar, H., Jablaoui, A., Gérard, P., et al. (2019). Microbial impact on cholesterol and bile acid metabolism: current status and future prospects. *J. Lipid Res.* 60, 323–332. doi: 10.1194/jlr.R088989

Labarthe, S., Polizzi, B., Phan, T., Goudon, T., Ribot, M., and Laroche, B. (2019). A mathematical model to investigate the key drivers of the biogeography of the colon microbiota. *J. Theor. Biol.* 462, 552–581. doi: 10.1016/j.jtbi.2018.12.009

Mc Auley, M. T., Wilkinson, D. J., Jones, J. J., and Kirkwood, T. B. (2012). A whole-body mathematical model of cholesterol metabolism and its age-associated dysregulation. *BMC Syst. Biol.* 6:130. doi: 10.1186/1752-0509-6-130

Mesmin, B., and Maxfield, F. R. (2009). Intracellular sterol dynamics. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1791, 636–645. doi: 10.1016/j.bbalip.2009.03.002

Millar, J. S., and Cuchel, M. (2018). Cholesterol metabolism in humans: a review of methods and comparison of results. *Curr. Opin. Lipidol.* 29, 1–9. doi: 10.1097/MOL.0000000000000475

Morgan, A. E., Mooney, K. M., Wilkinson, S. J., Pickles, N., and Mc Auley, M. T. (2016). Mathematically modelling the dynamics of cholesterol metabolism and ageing. *Biosystems* 145, 19–32. doi: 10.1016/j.biosystems.2016.05.001

Muñoz-Tamayo, R., Laroche, B., Walter, E., Doré, J., and Leclerc, M. (2010). Mathematical modelling of carbohydrate degradation by human colonic microbiota. *J. Theor. Biol.* 266, 189–201. doi: 10.1016/j.jtbi.2010.05.040

Pool, F., Currie, R., Sweby, P. K., Salazar, J. D., and Tindall, M. J. (2018). A mathematical model of the mevalonate cholesterol biosynthesis pathway. *J. Theor. Biol.* 443, 157–176. doi: 10.1016/j.jtbi.2017.12.023

Potiron, A., Gérard, P., Le Roy, T., Lesnik, P., Maguin, E., and Rhimi, M. (2015). Recent patents on hypocholesterolemic therapeutic strategies: an update. *Recent Adv. DNA Gene Seq.* 9, 36–44. doi: 10.2174/2352092210666151216143459

Read, M. N., and Holmes, A. J. (2017). Towards an integrative understanding of diet-host-gut microbiome interactions. *Front. Immunol.* 8:538. doi: 10.3389/fimmu.2017.00538

Ridlon, J. M., Kang, D. J., Hylemon, P. B., and Bajaj, J. S. (2014). Bile acids and the gut microbiome. *Curr. Opin. Gastroenterol.* 30:332. doi: 10.1097/MOG.0000000000000057

Russell, D. W. (2009). Fifty years of advances in bile acid synthesis and metabolism. *J. Lipid Res.* 50(Suppl.):S120–S125. doi: 10.1194/jlr.R800026-JLR200

Saltelli, A., Chan, K., and Scott, E. (2000). *Sensitivity Analysis*. Chichester: John Wiley and Sons.

Saltelli, A., Tarantola, S., and Chan, K.-S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* 41, 39–56. doi: 10.1080/00401706.1999.10485594

Schroepfer, G. J Jr. (2000). Oxysterols: modulators of cholesterol metabolism and other processes. *Physiol. Rev.* 80, 361–554. doi: 10.1152/physrev.2000.80.1.361

Sekimoto, H., Shimada, O., Makanishi, M., Nakano, T., and Katayama, O. (1983). Interrelationship between serum and fecal sterols. *Jpn. J. Med.* 22, 14–20. doi: 10.2169/internalmedicine1962.22.14

Sips, F. L., Tiemann, C. A., Oosterveer, M. H., Groen, A. K., Hilbers, P. A., and van Riel, N. A. (2014). A computational model for the analysis of lipoprotein distributions in the mouse: translating FPLC profiles to lipoprotein metabolism. *PLoS Comput. Biol.* 10:e1003579. doi: 10.1371/journal.pcbi.1003579

Swann, J. R., Want, E. J., Geier, F. M., Spagou, K., Wilson, I. D., Sidaway, J. E., et al. (2011). Systemic gut microbial modulation of bile acid metabolism in host tissue compartments. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4523–4530. doi: 10.1073/pnas.1006734107

Tabas, I. (2002). Cholesterol in health and disease. *J. Clin. Investig.* 110, 583–590. doi: 10.1172/JCI0216381

Tanaka, H., Doesburg, K., Iwasaki, T., and Mierau, I. (1999). Screening of lactic acid bacteria for bile salt hydrolase activity. *J. Dairy Sci.* 82, 2530–2535. doi: 10.3168/jds.S0022-0302(99)75506-2

Thompson, G., O'Neill, F., and Seed, M. (2002). Why some patients respond poorly to statins and how this might be remedied. *Eur. Heart J.* 23, 200–206. doi: 10.1053/euhj.2001.3071

van de Pas, N. C., Soffers, A. E., Freidig, A. P., van Ommen, B., Woutersen, R. A., Rietjens, I. M., et al. (2010). Systematic construction of a conceptual minimal model of plasma cholesterol levels based on knockout mouse phenotypes. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1801, 646–654. doi: 10.1016/j.bbalip.2010.02.009

van de Pas, N. C., Woutersen, R. A., van Ommen, B., Rietjens, I. M., and de Graaf, A. A. (2011). A physiologically-based kinetic model for the prediction of plasma cholesterol concentrations in the mouse. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1811, 333–342. doi: 10.1016/j.bbalip.2011.02.002

van de Pas, N. C., Woutersen, R. A., van Ommen, B., Rietjens, I. M., and de Graaf, A. A. (2012). A physiologically based *in silico* kinetic model predicting plasma cholesterol concentrations in humans. *J. Lipid Res.* 53, 2734–2746 doi: 10.1194/jlr.M031930

Van der Velde, A. E., Vrins, C. L., Van den Oever, K., Kunne, C., Elferink, R. P. O., Kuipers, F., et al. (2007). Direct intestinal cholesterol secretion contributes significantly to total fecal neutral sterol excretion in mice. *Gastroenterology* 133, 967–975. doi: 10.1053/j.gastro.2007.06.019

Wang, D. Q., and Carey, M. C. (2003). Measurement of intestinal cholesterol absorption by plasma and fecal dual-isotope ratio, mass balance, and lymph fistula methods in the mouse an analysis of direct versus indirect methodologies. *J. Lipid Res.* 44, 1042–1059. doi: 10.1194/jlr.D200041-JLR200

Wang, D. Q., Paigen, B., and Carey, M. C. (2001). Genetic factors at the enterocyte level account for variations in intestinal cholesterol absorption efficiency among inbred strains of mice. *J. Lipid Res.* 42, 1820–1830. Available online at: https://www.jlr.org/content/42/11/1820.short

World Health Organization (2017). Cardiovascular diseases (CVDs). Available online at: http://www.who.int/cardiovascular_diseases/en/

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 12 June 2020
doi: 10.3389/fmicb.2020.01106

Gut Microbiota-Based Algorithms in the Prediction of Metachronous Adenoma in Colorectal Cancer Patients Following Surgery

*Yang Liu†, Rui Geng†, Lujia Liu, Xiangren Jin, Wei Yan, Fuya Zhao, Shuang Wang, Xiao Guo, Ghanashyam Ghimire and Yunwei Wei**

Department of Oncological and Endoscopic Surgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China

Evaluating the risk of colorectal metachronous adenoma (MA), which is a precancerous lesion, is necessary for metachronous colorectal cancer (CRC) precaution among CRC patients who had underwent surgical removal of their primary tumor. Here, discovery cohort ($n = 41$) and validation cohort ($n = 45$) of CRC patients were prospectively enrolled in this study. Mucosal and fecal samples were used for gut microbiota analysis by sequencing the 16S rRNA genes. Significant reduction of microbial diversity was noted in MA ($P < 0.001$). A signature defined by decreased abundance of eight genera and increased abundance of two genera strongly correlated with MA. The microbiota-based random forest (RF) model, established utilizing *Escherichia–Shigella*, *Acinetobacter* together with BMI in combination, achieved AUC values of 0.885 and 0.832 for MA, predicting in discovery and validation cohort, respectively. The RF model was performed as well for fecal and tumor adjacent mucosal samples with an AUC of 0.835 and 0.889, respectively. Gut microbiota profile of MA still existed in post-operative cohort patients, but the RF model could not be performed well on this cohort, with an AUC of 0.61. Finally, we introduced a risk score based on *Escherichia–Shigella*, *Acinetobacter* and BMI, and synchronous-adenoma achieved AUC values of 0.94 and 0.835 in discovery and validation cohort, respectively. This study presented a comprehensive landscape of gut microbiota in MA, demonstrated that the gut microbiota-based models and scoring system achieved good ability to predict the risk for developing MA after surgical resection. Our study suggests that gut microbiota is a potential predictive biomarker for MA.

Keywords: colorectal cancer, metachronous cancer, colorectal adenoma, gut microbiota, random forest

OPEN ACCESS

Edited by:
Steve Lindemann, Purdue University, United States

Reviewed by:
Francesco Vitali, National Research Council (CNR), Italy
Jonathan Badger, National Cancer Institute (NCI), United States

***Correspondence:**
Yunwei Wei
hydwyw11@hotmail.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *02 March 2020*
Accepted: *04 May 2020*
Published: *12 June 2020*

Citation:
Liu Y, Geng R, Liu L, Jin X, Yan W, Zhao F, Wang S, Guo X, Ghimire G and Wei Y (2020) Gut Microbiota-Based Algorithms in the Prediction of Metachronous Adenoma in Colorectal Cancer Patients Following Surgery. Front. Microbiol. 11:1106. doi: 10.3389/fmicb.2020.01106

INTRODUCTION

Colorectal cancer (CRC) is among the leading cause of cancer-related deaths worldwide. Despite substantial progress in the early diagnosis and treatment of CRC and the fact that more than two-thirds of CRC patients received surgical resection and adjuvant therapy, nearly 40% of these patients developed CRC recurrence, including local recurrence, metachronous cancer, and distant metastasis (Kahi et al., 2016). It has been well-documented that patients with a history of CRC are at an increased risk of developing metachronous CRC following surgical resection and pre-operative clearing (Balleste et al., 2007; Mulder et al., 2012). As such, post-operative colonoscopy is

highly recommend for patients after surgical resection of CRC to improve survival via diagnosing metachronous CRC at an early stage or to prevent the occurrence of metachronous CRC via detecting and removing of the pre-cancerous colorectal polyps (Kahi et al., 2016). According to the major guidelines, an initial full colonoscopy is recommended at the time of diagnosis or within 3–6 months following surgical intervention for detection of synchronous lesions, while further colonoscopies should be carried out >6 months, generally 1 year after the surgical resection (Meyerhardt et al., 2013), followed by colonoscopies every 3–5 years for detection of metachronous cancer. There is no first-level evidence in support of the optimal total duration of surveillance after treatment for CRC (van der Stok et al., 2016).

Despite a certain level risk, there has been a lack of reliable factors to be used for predicting metachronous CRC in patients who have undergone surgical treatment. Thus, life-long colonoscopy surveillance is needed. Currently, several factors have been shown to be associated with an increased risk of metachronous CRC, including age, previous or synchronous adenomas or history of CRC, right-sided tumors, and microsatellite instability (MSI); many of these reported risk factors were inconsistent in the previous studies (Balleste et al., 2007; Bouvier et al., 2008). Identification of individuals at high risk for the development of metachronous colorectal cancer is necessary to increase the efficiency of surveillance and to improve prognosis.

Recent studies have suggested that the community of microbes inhabiting the gastrointestinal tract plays an important role in the development and progression of CRC (Arthur et al., 2012; Kostic et al., 2013). In fact, gut microbiota dysbiosis was already found in patients with colorectal adenoma, and the disturbance became more apparent during the progression of adenoma into CRC (Feng et al., 2015). It has been of note that gut bacteria may exert a role in tumorigenesis, and in turn, they may have potential as useful biomarkers for the early detection of disease (Zeller et al., 2014). A previous study has indicated that gut microbiota could be used to quantify the risk of recurrence (Sze et al., 2017). Until now, it remains unknown if gut microbiota could hold a value in assessment of risk for metachronous CRC or precancerous lesions such as colorectal adenoma, given the pathogenesis of CRC.

As CRC develops gradually from premalignant adenomatous, accurate prediction and early detection polyps provides an opportunity to halt this process. Our previous study found that colorectal cancer patients who developed metachronous adenoma (MA) post-operatively showed distinct fecal microbiota, which can be potentially used for diagnosis for MA (Jin et al., 2019). But, the features of MA gut microbiota that already existed before operation or formed post-operatively is still unknown. Could pre-operative gut microbiota be used as a tool to predict the risk for post-operative MA?

In this study, discovery and validation cohort of CRC patients was prospectively enrolled, the mucosal and fecal samples were used for analysis of gut microbiota by sequencing the 16S rRNA genes. We aimed to test the hypothesis that the gut microbiota composition before surgery was associated with the risk of developing MA and thus could be used, together with other independent risk factors, to generate new algorithms for better predicting MA.

MATERIALS AND METHODS

Study Population

Colorectal cancer patients of discovery and validation cohort were both prospectively enrolled at the First Affiliated Hospital of Harbin Medical University during the period between September 2017 and April 2018. All the patients were diagnosed with primary colorectal adenocarcinoma and underwent surgical resection of CRC. During the enrollment, the patients who had the following conditions were excluded from this study: (1) taking antibiotics in 1 month prior to colonoscopy examination; (2) previous diagnosis of CRC, IBD, or IBS; and (3) medical history of surgery, radiation, or chemotherapy. A total of 41 CRC patients were in the discovery cohort, 13 patients had both fecal and colonoscopic mucosal samples, and 45 patients in the validation cohort had colonoscopic mucosal samples, but not fecal samples.

Sample Collection

Cold biopsy forceps were used for collection of colonoscopic mucosal biopsies from CRC tissues and adjacent, cancer-free tissues (at least 5 cm away from lesions), respectively. Fecal samples were taken the night before colonoscopy examination day. All the samples were snap-frozen in cryovial immediately following collection and stored at −80°C until DNA extraction.

Follow-Up

All the study patients were followed up for 12 months; during the follow-up period, they were scheduled to undergo surveillance colonoscopy every 1 year. Four patients in the discovery cohort and seven patients in the validation cohort, combined with a malignant bowel obstruction (MBO), were asked for colonoscopy within 2–4 months after surgery to detect synchronous lesions, followed by repeat colonoscopy at 1 year to detect metachronous lesion according to the guidelines (Meyerhardt et al., 2013). For patients with synchronous adenoma detected before surgery, endoscopic mucosa resection (EMR) was performed to remove the lesion prior to colon resection. The primary endpoint was MA detection during follow up period.

DNA Extraction and 16S rRNA Gene Sequencing for Bacterial Identification

The fecal and mucosal samples as described in the sample collection were used for DNA extraction. In brief, microbial DNA was extracted using a DNA kit (Bio-Tek, GA, United States) according to the manufacturer's instructions and used for an amplification of the hypervariable regions (V3-V4) of the bacterial 16S rRNA gene. The resulting amplicons were purified and pooled in equimolar concentrations, followed by paired-end sequencing (2 × 300) on an Illumina MiSeq platform (Illumina, San Diego, CA, United States), which was performed by Majorbio Bio-Pharm Technology (Shanghai, China). After the raw reads

were filtered and quality control was conducted, OTUs were clustered with a 97% similarity cut-off using UPARSE[1] (version 7.1), following which, the identified chimeric sequences were removed using UCHIME. With the RDP Classifier algorithm, taxonomic assignments for the 16S rRNA gene sequences were made[2] with the GreenGene 16S rRNA gene database at a confidence threshold of 70%. The 16S rRNA gene sequencing runs were separately performed for the discovery and validation cohorts for both MA and nMA patients.

Bioinformatics and Statistical Analysis

Both α-diversity (Simpson-reciprocal and Shannon indices) and β-diversity (Bray–Curtis distance) were examined using QIIME (Version 1.7.0). PCoA was used to reduce the dimension of the original variables with the Vegan and ggplot2 packages in R, while Analysis of similarity (ANOSIM) of the distance matrices in the vegan package in R was used to quantize the similarity and test the statistical significance between groups (Buttigieg and Ramette, 2014). Hierarchical clustering on the basis of similarities in the combination of variables was carried out using Pvclust in R. The microbiota were characterized using the linear discriminant analysis effect size (LEfSe) method for representative taxa discovery, emphasizing both significance and biological relevance (Segata et al., 2011). Functional composition of the gut metagenomes were predicted and profiled in accordance with the 16S rRNA gene sequences using PICRUSt with level III KEGG database pathways (Langille et al., 2013). Both PICRUSt and LEfSe were accomplished online[3]. A heatmap was created to express the results with the heatmap package in R. The microbiota features were further analyzed as categorical variables using an univariate logistic regression to screen risk factors. The optimal cut-off for each bacterial group was determined by ROC analysis. Variables with a P value < 0.1 on the univariate analysis were selected for further forward stepwise multivariate logistic regression to identify independent predictors. Odds ratios (ORs) were calculated with a 95% confidence interval (CI). The random forest (RF) algorithm was used to create the classification models. The optimal number of variables was determined by maximizing the area under the curve of the receiver operator characteristic (AUC) with the AUCRF package, then caret (v6.0.76) and random forest R package were used to build model. To avoid over-fitting of the data in the model, 10-time and 10-fold cross-validations were made. The resulting model was subsequently used for validation cohort.

All categorical data were presented as number of cases and percentages, while continuous data were shown as median with range. Categorical variables were compared by the Pearson's chi-square (χ^2) test, and continuous variables by Mann–Whitney U test where appropriate. Statistical analysis of the data was performed using SPSS (SPSS version 19, La Jolla, CA, United States). Wilcoxon rank sum test and Multiple hypothesis tests were used for analysis of continuous and categorical data and adjusted using the Benjamini and Hochberg FDR. The results with an FDR threshold lower than 0.1 were considered significant differences. Spearman's rank test was used for correlation analysis, and a P value less than 0.05 was considered statistically significant.

[1]http://drive5.com/uparse/

[2]http://rdp.cme.msu.edu/

[3]http://huttenhower.sph.harvard.edu/galaxy

RESULTS

Characteristics of the Study Patients

Forty-one patients were included for discovery cohort, of which 22 patients developed metachronous adenoma (MA group), and the remaining 19 patients did not have any signs of metachronous adenoma [non-metachronous adenoma (nMA) group]. Demographic and clinical features between the two groups were summarized in **Table 1**. Body mass index (BMI) in the MA group was significantly greater than that of the nMA group (25.25 vs. 23.0; $P < 0.05$). Notably, the incidence of synchronous adenoma was significantly higher in the MA versus nMA groups (15/22 vs. 7/19; $P < 0.05$). No other significant differences between the two groups were observed. Information for every participant were supplied in **Supplementary Table S1**. Another 45 patients were included for validation cohort, 21 of which developed MA (**Supplementary Table S2**).

Mucosal Microbial Diversity Is Significantly Associated With Metachronous Adenoma

We initially examined the correlation between mucosal microbial diversity and the development of MA. As shown in **Supplementary Figure S1**, the 16S rRNA gene-sequencing reads and depths were adequate. An analysis of the mucosal microbial diversity with two methodologies (Shannon and Simpson-reciprocal indices) showed that alpha-diversity of the mucosal microbiome was significantly higher in the nMA group compared with the MA group ($P < 0.001$ for each index) (**Figures 1A,B**). A principal coordinate analysis (PCoA) on genus level with Bray–Curtis metric distance was performed for comparison of β-diversity between the two groups. As shown in **Figure 1C**, a clear clustering between the MA and nMA groups was revealed, suggesting that the mucosa microbial communities exhibited phylogenetic closeness within each group ($P = 0.001$). Importantly, we excluded the possibility of any other potential contributors to the microbial diversity, such as clinical-pathological features, synchronous adenoma, BMI, sex, and adjuvant therapies (**Supplementary Figure S2**).

Mucosal Microbial Composition and Function in the MA Group Differs Significantly From Those in the nMA Group

We next determined if there were differences in the mucosal microbial composition between the MA and nMA patients using linear discriminant analysis of effect size (LEfSe). After bacterial taxa with relative abundance <0.5% were excluded for comparison, 10 taxa showed differentiated distribution with LDA score > 4.0 on genus level. The MA group exhibited a predominance of *Escherichia–Shigella* and *Roseburia*, while the nMA group had a predominance of

Prevotella_9, *Herbaspirillum*, *unclassified_k_norank_d_Bacteria*, *Acinetobacter*, *Blautia*, *Faecalibacterium*, *Rhodococcus*, and *Ruminococcus_torques_group* (**Figure 1D**). We then examined the potential interactions among these 10 taxa with Spearman rank test. As a result, *Escherichia–Shigella* was always negatively correlated (red dots) with others taxa, while the genera enriched in the nMA group (green text) positively correlated (blue dots) with each other (**Figure 1E**).

Further analysis showed there were four taxa on the phylum level and six taxa on the family level that predominated in the two groups with LDA score > 4.0 (**Supplementary Figure S3**). We then interrogated whether the mucosal microbiome can be segregated using BMI or synchronous adenoma as grouping variables. Only one and two predominate genera with LDA score > 4.0 were found, respectively, based on BMI (high or normal) and synchronous adenoma status (**Supplementary Figure S4**), indicating that MA rather than BMI or synchronous adenoma was the main explanation to the different microbiota composition between the two groups.

TABLE 1 | Clinico-pathological characteristics of patients.

	MA (n = 22)	nMA (n = 19)	P value
Gender			
Female	12	6	0.139
Male	10	13	
Age (years)[a]	63 (58.5–68.75)	61.3 (53–68.5)	0.619
BMI[a]	25.25 (22.75–27.98)	23.0 (21.74–23.7)	0.011*
Synchronous adenoma			
Yes	15	7	0.045*
No	7	12	
Bowel obstruction[d]			
Yes	2	2	0.877
No	20	17	
Hematochezia			
Yes	11	11	0.613
No	11	8	
Tumor size[ac]	4 (3.6–4.2)	4 (3.1–4.75)	0.854
Tumor location[b]			
Left hemi-colon	7	2	0.171
Right hemi-colon	3	6	
Rectum	12	11	
CEA[a]	6.725 (2.38–14.30)	3.97 (2.37–12.83)	0.896
CA 19-9[a]	12.31 (7.15–65.44)	12.55 (10.99–20.06)	0.744
Adjuvant therapy			
Yes	13	12	0.790
No	9	7	
TNM-stage			
I	2	2	0.537
IIA	17	11	
IIIA	0	1	
IIIB	3	5	

[a]Data shown as median (1st and 3rd quartile). [b]Tumor location: splenic flexure, descending, sigmoid, rectosigmoid were classified as left hemi-colon; ileocecal, ascending, hepatic flexure, transverse were classified as right hemi-colon. [c]Tumor size definition: maximum diameter. [d]Bowel obstruction was defined when coloscopy cannot pass through the tumor obstruction.

The functions of the gut microbiota were predicted using the PICRUSt analysis. 16S rRNA gene sequencing data were categorized into 328 KEGG functional pathways; pathways present in <10% of participants were removed, leaving 284 KEGG pathways for comparation. Fifty five pathways were differentially enriched between the two groups (P_{fdr} < 0.1) (**Supplementary Figure S5**). We observed significant upregulation of *bacterial invasion of epithelial cells pathway* and *lipopolysaccharide biosynthesis protein pathway* in the MA group compared with the nMA group (P_{fdr} < 0.1). On the contrary, *p53 signal pathway* was downregulated in the MA group (P_{fdr} < 0.1) (**Figures 1F–H**). Specifically, the potential pathogenic bacteria *Escherichia–Shigella* was positively correlated with *bacterial invasion of epithelial cells pathway* (r = 0.89, P < 0.01) (**Figure 1I**).

Microbiota Profiles of the Mucosal and Fecal Samples

Bar plots of the class taxonomic levels showed Gammaproteobacteria and Clostridia as the top two classes with higher relative abundance in all samples. *P < 0.05, different from controls by Wilcoxon rank-sum test or Chi-squared test for continuous or categorical variables, respectively. The microbiota composition was similar between on-tumor and off-tumor mucosal samples, whereas fecal samples showed independent features without detecting of *unclassified_k__norank_d__Bacteria* and *Fusobacteriia* (**Figure 2A**). Despite the collective differences between subjects with MA and nMA, the microbiota associated with on-tumor and off-tumor tissues in the same individual (n = 12) did not differ significantly in PCoA (**Figure 2B**) (P = 0.691). Hierarchical-Clustering analysis with Bray–Curtis distance indicated no apparent difference between the paired On/Off mucosal samples in the same individual (**Supplementary Figure S6**). On the contrary, fecal and mucosal samples in the same individual showed obviously different in PCoA (**Figure 2C**) (P = 0.001), paired fecal and mucosa samples within the same individual did not close to each other (**Supplementary Figure S7**).

Next, we assessed whether fecal microbiota profiles could reflect the difference between MA (n = 11) and nMA (n = 8). As expected, fecal microbiota profiles in the MA and nMA patients differed significantly in PCoA analysis (**Supplementary Table S3** and **Supplementary Figure S8**) (P = 0.003). The microbiota of the fecal samples in LEfSe analysis by MA status produced five genera with LDA score > 4.0, with *Escherichia–Shigella*, *Blautia*, and *Ruminococcus_torques_group* profiles consistent with the findings of the mucosal profiling (**Supplementary Figure S9**). These results indicated that even though fecal microbiota do not corresponded to mucosa microbiota and only partially reflect the microbiota at the mucus layer, differences due to disease status are still evident.

Gut Microbiota Variation of MA May Still Exist to Some Degree in Patients After Surgery

Our previous cross-sectional study showed significant difference in post-operative fecal microbiota between patients with and

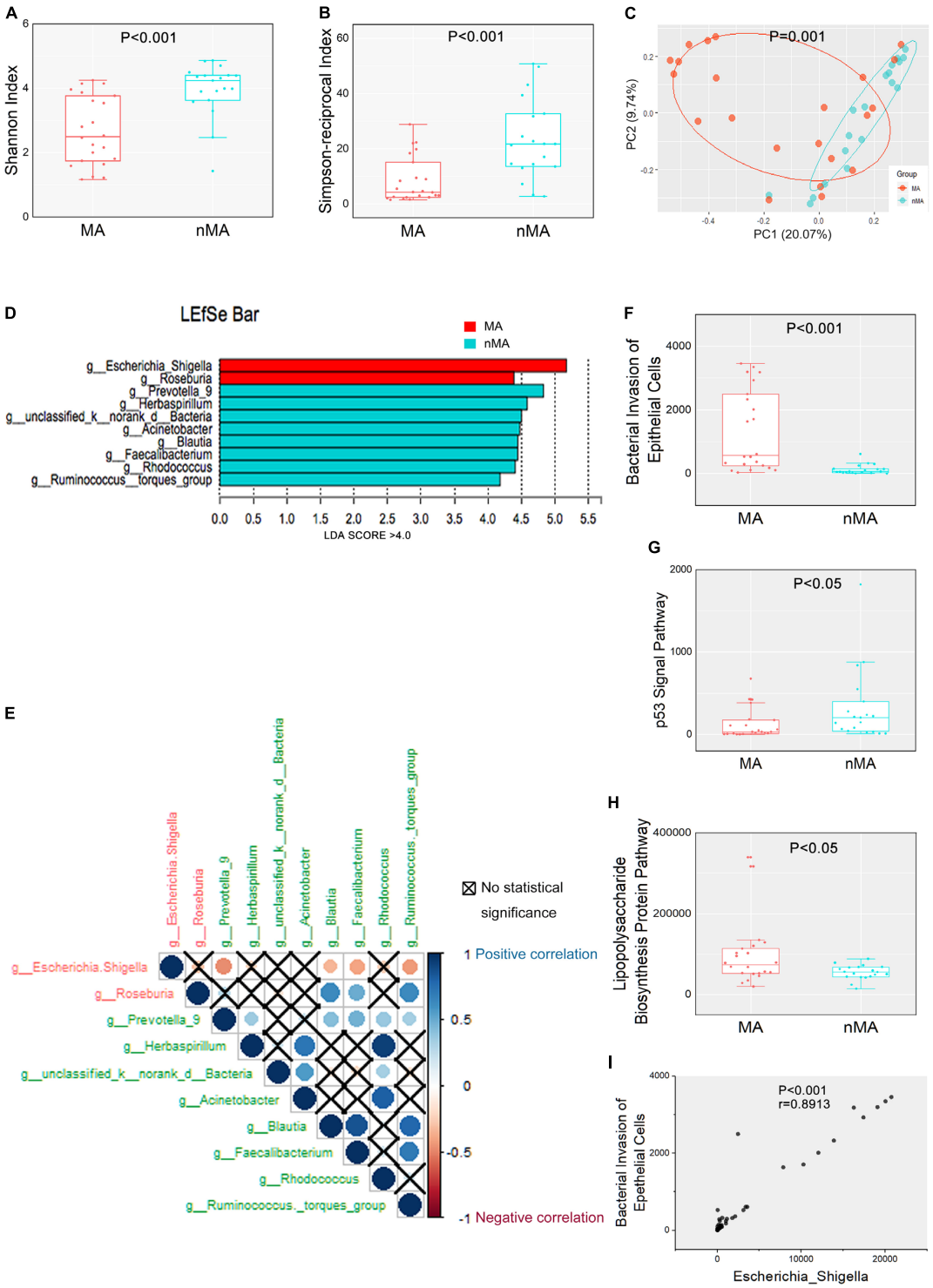


FIGURE 1 | Mucosal microbiome diversity and communities are significantly different between MA and nMA. **(A,B)** α-diversity boxplot (Shannon and Simpson-reciprocal indices) of mucosal samples in MA and nMA groups. Boxes represented the 25th to 75th percentile of the distribution; the median was shown as a thick line in the middle of the box; whiskers extend to values with 1.5-times the difference between the 25th and 75th percentiles. **(C)** PCoA using Bray–Curtis of β-diversity in MA and nMA groups. **(D)** LDA score computed from features differentially abundant between MA and nMA in mucosal samples. The criteria for feature selection was log LDA score > 4. **(E)** Spearman correlations among two MA-enriched (red) and eight-nMA enriched (green) genera taxa in mucosal samples of CRC patients. Red dots indicated negative correlation, blue dots indicated positive correlation, cross indicated no significance ($P > 0.05$). **(F–H)** Boxplot of *bacterial invasion of epithelial cells pathway*, *Lipopolysaccharide biosynthesis protein pathway*, and *p53 signal pathway* between MA and nMA. *P* values were adjusted using the FDR correction. **(I)** Spearman correlation between *bacterial invasion of epithelial cells pathway* and relative abundance of *Escherichia–Shigella*.

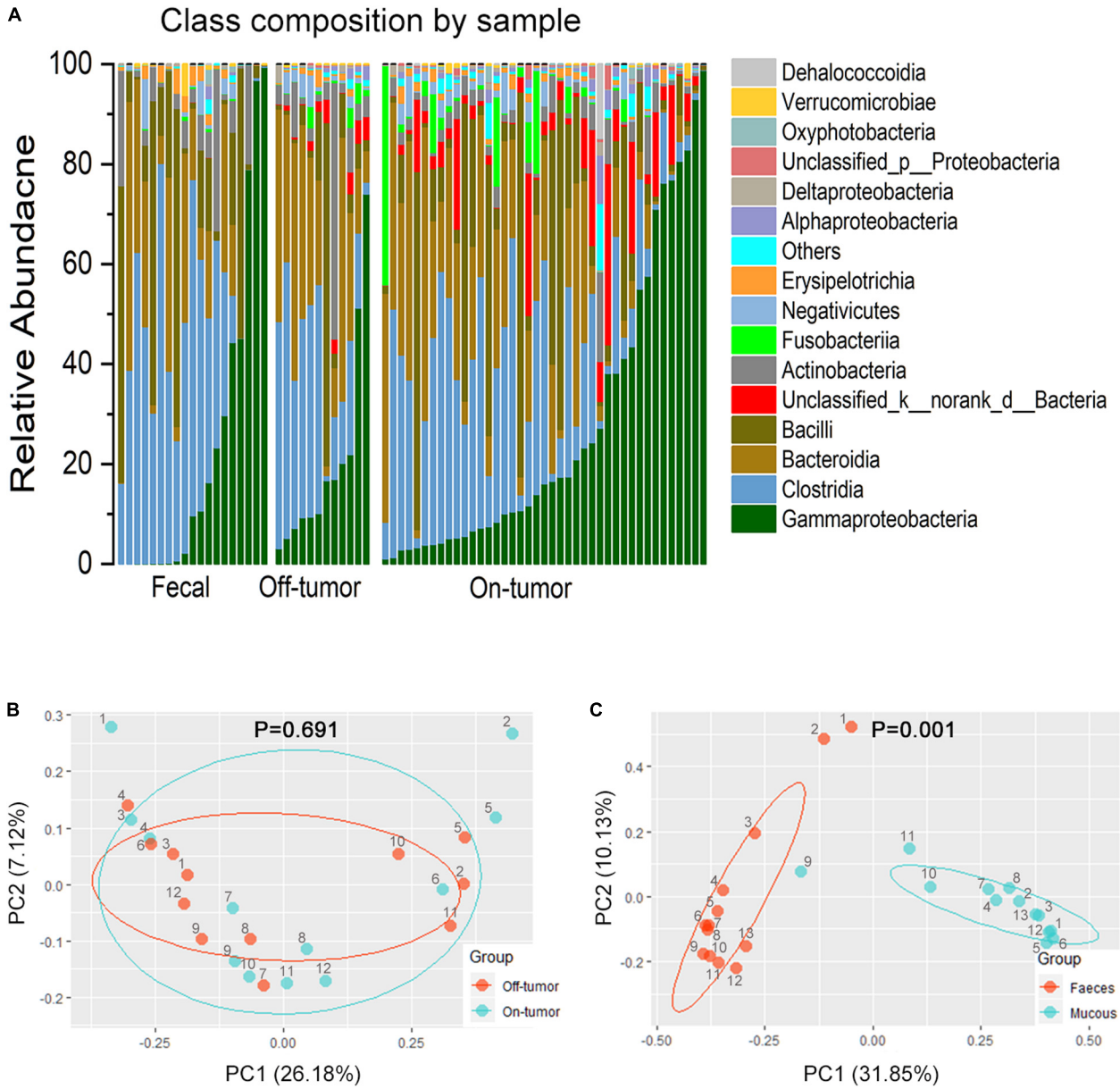


FIGURE 2 | Fecal and off-tumor samples. **(A)** Bar plots of the class taxonomic levels of microbiota in fecal, off-tumor and on-tumor samples. Relative abundance is plotted for each samples. **(B)** PCoA using Bray–Curtis of β-diversity between on- and off-tumor mucosal samples. **(C)** PCoA using Bray–Curtis of β-diversity between fecal and mucosal samples.

without MA, and the alterations in the gut microbiota was associated with the disease progression in health-adenoma-carcinoma sequence (Feng et al., 2015), indicating that patients with occurrence of metachronous had more "carcinoma-like" gut microbiota compared to clear-intestine patients. Intrigued by these pervious findings, we examined if there was an association between pre- and post-operative patients fecal microbiota on MA profile. To this end, we applied conjoint analysis by importing our previous 16S rRNA gene sequence data of fecal samples, assigned as post-operative cohort. The samples from this study were assigned as pre-operative cohort accordingly.

The overall α-diversity of post-operative patients ($n = 47$) was higher than that of pre-operative patients ($n = 19$) (data not shown). Similarly, α-diversity of the fecal samples were higher in the nMA patients ($P < 0.05$ for both Shannon and Simpson-reciprocal indices). For post-operative patients, α-diversity was higher in the nMA patients, whereas the difference was not statistically significant ($P > 0.05$ for both Shannon and Simpson-reciprocal indices) (**Figures 3A,B**). Next, *Escherichia–Shigella* was selected, as it was highly enriched and relatively abundant in both the mucosal and fecal samples in the MA patients ($P < 0.05$). In addition, this difference was also found in post-operative

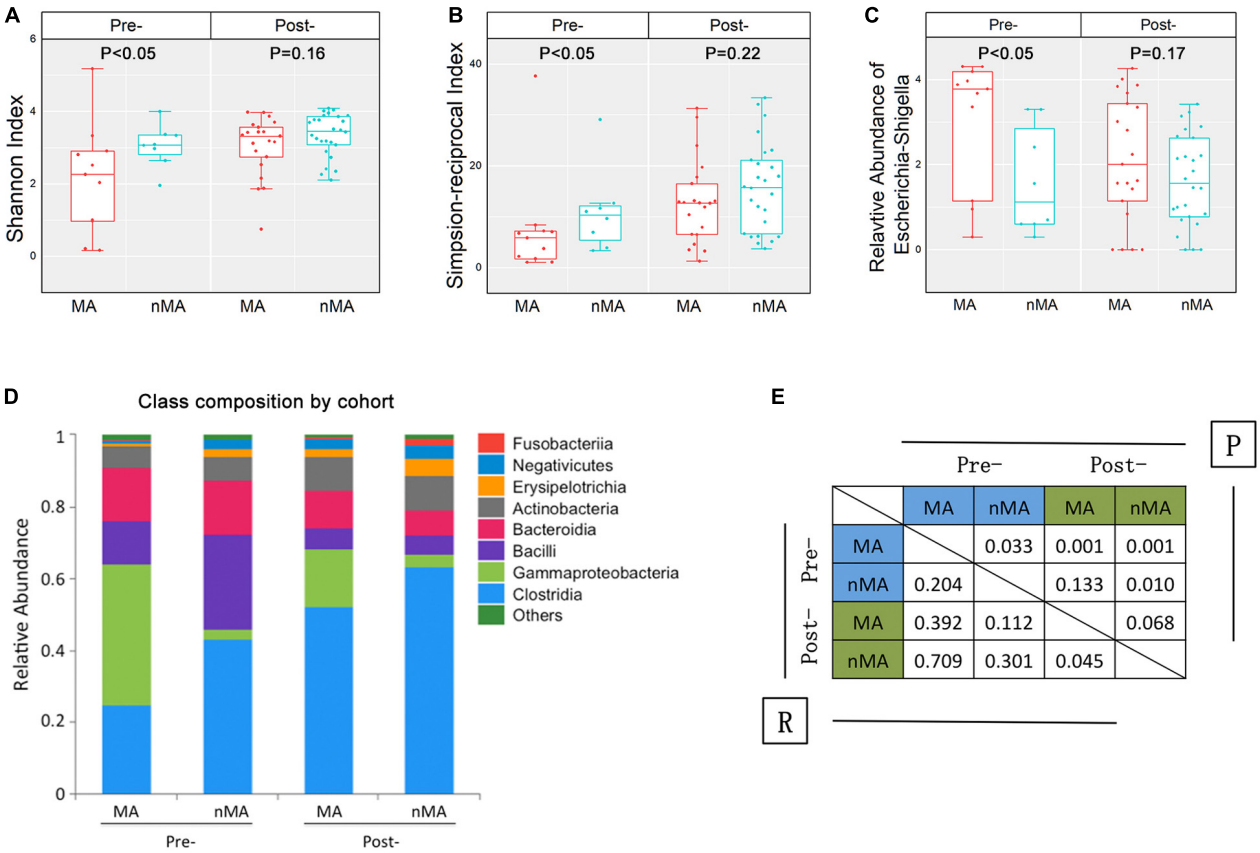


FIGURE 3 | Fecal microbiota in CRC patients and CRC patients after surgical therapy. **(A,B)** α-diversity boxplot (Shannon and Simpson-reciprocal indices) of fecal samples. **(C)** Boxplots of relative abundance of fecal *Escherichia–Shigella*; boxplot illustration was provided in **Figure 1**. **(D)** Bar plots of the class taxonomic levels of fecal microbiota. Relative abundance is plotted for each group. **(E)** ANOSIM result between fecal samples of groups. *R* value indicated the strength of the factors on the samples, while give *P*-value indicated the significance levels.

patients without reaching statistical significance (**Figure 3C**). Bar plots of the class taxonomic levels showed a difference in the microbiota composition between the MA and nMA patients, as well as between the post-MA and post-nMA patients. It was worth noticing that the microbiota composition of the MA patients was similar to that of the post-MA patients, while that of nMA was more similar to post-nMA (**Figure 3D**).

ANOSIM was performed to determine the β-diversity between groups, in which ANOSIM gave a *P* value (i.e., significance levels) and a *R* value (i.e., the strength of the factors on the samples). As a result, the *R* value between the MA and nMA groups was 0.204 ($P = 0.033$), while *R* value between the post-MA and post-nMA groups was 0.045 ($P = 0.068$), indicating that the discrepancy between patients with and without MA was less obvious in patients undergone surgery compared to untreated patients. *R* values between post-nMA and MA or nMA ($R = 0.709$ or $R = 0.301$; $P = 0.001$ or $P = 0.01$) were higher than those between post-MA and MA or nMA ($R = 0.392$ or $R = 0.112$; $P = 0.001$ or $P = 0.133$) (**Figure 3E**), suggesting that gut microbiota of post-operative patients without MA to be more different from CRC patients, especially from CRC patients who develop MA. Collectively, these results indicated that gut microbiota-based discrepancy between patients with and without MA remained in post-operative patients.

Pre-operative Gut Microbiota-Based Random Forest Algorithms and Scoring System in the Prediction of Metachronous Adenoma in CRC Patients After Surgery

Firstly, 7 of 10 predominance bacterial genera in MA and nMA identified by LEfSe analysis, together with BMI, and synchronous adenoma were applied to logistic regression. *Herbaspirillum*, *Rhodococcus,* and *Prevotella_9* were excluded, as they were not detectable in more than five patients. All these variables were identified as significant risk factors for MA by univariate logistic regression ($P \leq 0.1$) (**Table 2**), then multivariate logistic regression analysis was applied for independent risk factor validation. As shown in **Table 3**, the predominant bacterial genera, including *Escherichia–Shigella* and *Acinetobacter,* as well as BMI were identified as independent risk factors for MA ($P < 0.05$), with a good ability for differentiating MA from nMA (AUC, 0.935).

TABLE 2 | Univariate logistic regression predicting MA.

	Cut-off value	OR	95% CI	*P* value
Escherichia–Shigella	564.5	10.000	2.350–42.547	0.002*
unclassified_k_norank_d_Bacteria	147	0.206	0.037–1.131	0.069
Faecalibacterium	608.5	0.172	0.044–0.672	0.011*
Ruminococcus_torques_group	10.5	0.097	0.011–0.871	0.037*
Blautia	732.5	0.065	0.007–0.593	0.015*
Acinetobacter	28	0.056	0.006–0.492	0.009*
Roseburia	55	0.172	0.044–0.672	0.011*
Synchronous adenoma		3.673	1.007–13.395	0.049*
BMI		1.396	1.069–1.824	0.014*

TABLE 3 | Multivariable logistic regression model predicting MA.

	OR	95% CI	*P* value
Escherichia–Shigella	53.254	3.338–849.676	0.005*
Acinetobacter	0.026	0.001–0.477	0.014*
BMI	1.684	0.993–2.855	0.053

Next, we constructed an RF algorithm using the relative abundance of the gut microbial populations with or without the clinical risk factors to predict MA. To determine the potential of bacterial taxa in discriminating MA, we aimed to identified a minimal set of bacterial genera that maximally differentiated nMA from MA. Firstly, 10 predominant bacterial genera produced by LEfSe were initially screened, and a combination of *Escherichia–Shigella* and *Acinetobacter* optimized the performance of RF model (**Supplementary Figure S10**), and thus were used to generate a new model. 10-times and 10-fold cross-validations were conducted to optimize the model in case of over-fitting. As shown in **Figure 4**, the AUC for the model was 0.809 and higher than *Escherichia–Shigella* or *Acinetobacter* alone in predicting MA (**Figure 4A**). Considering the potential value of some clinical factors in the prediction of MA, we hypothesized that the predominant bacterial populations and clinical factors in combination could generate a more precise RF model. To test the hypothesis, the independent clinical risk factors, including synchronous adenoma and BMI (**Supplementary Figure S11**), together with the predominant bacterial populations, *Escherichia–Shigella* and *Acinetobacter,* were used to build a new RF model. The AUC for the RF model was 0.885, which was greater than the AUC for the RF model using predominant bacterial populations alone (**Figure 4A**). This result indicated that, in addition to gut microbiota, clinical features of patients possessed additional predictive ability on MA. The RF model were further tested on fecal and off-tumor samples, the AUC was 0.835 and 0.889, respectively (**Supplementary Figures S12**, **S13**), suggesting that fecal and off-tumor mucosal samples can be used for MA prediction as well. However, the AUC for the RF model was 0.61 on post-operative fecal samples (**Supplementary Figure S14**). Finally, the RF model was applied for discovery cohort and got a AUC of 0.832 (**Figure 4B**).

In order to further validated the specificity of our RF model, we applied the RF model to predict local recurrence of colon cancer with previous published data (Bullman et al., 2017). The AUC value was 0.546, which indicated a poor predict ability for local recurrence (**Supplementary Figure S15**).

Finally, we developed a risk score for MA, which utilized the two predominant bacterial populations and the two clinical features. *Escherichia–Shigella*, BMI and synchronous-adenoma were risk factors, and the presence of each one was assigned one point, while the absence of beneficial factor, *Acinetobacter,* was scored one point. The cut-off values were determined by ROC analysis in the discovery cohort and applied the same value for the validation cohort to avoid over-fitting. As a result, the total risk scores ranged from zero to four points, and the risk score showed an AUC of 0.94 and 0.835 for the prediction of MA in discovery and validation cohort. Further, the presence of two or more risk factors in discovery cohort had a sensitivity and specificity of 90.9% and 89.5%, but specificity in validation cohort was 33.3% (**Table 4**).

DISCUSSION

We conducted the first study, to the best of our knowledge, to assess the correlation between pre-operative gut microbiota and MA among Chinese CRC patients after surgery and to develop novel microbiota-based predictive models. The novel findings are summarized as follows: (1) There was a significant correlation between pre-operative gut microbiota and the development of MA among CRC patients after surgery. (2) Specific members of the predominant gut microbiota, including *Escherichia–Shigella* and *Acinetobacter,* were identified as independent risk factors for MA. (3) The microbiota-based RF model was established utilizing these specific members of predominant gut microbiota combined with independent clinical risk factors (BMI) and the status of synchronous adenoma, showing a good performance (AUC, 0.885) to predict MA among CRC patients after surgery. (4) The microbiota-based RF model exhibited good ability in the prediction of MA using fecal and off-tumor samples (AUC, 0.835 and 0.889, respectively). (5) A risk-scoring system was proposed with four independent predictive factors got an AUC of 0.94 and 0.835 for the prediction of MA in discovery and validation cohort.

Colonoscopic mucosa biopsies were used rather than an intra-operative specimen, because we thought the microbiota of

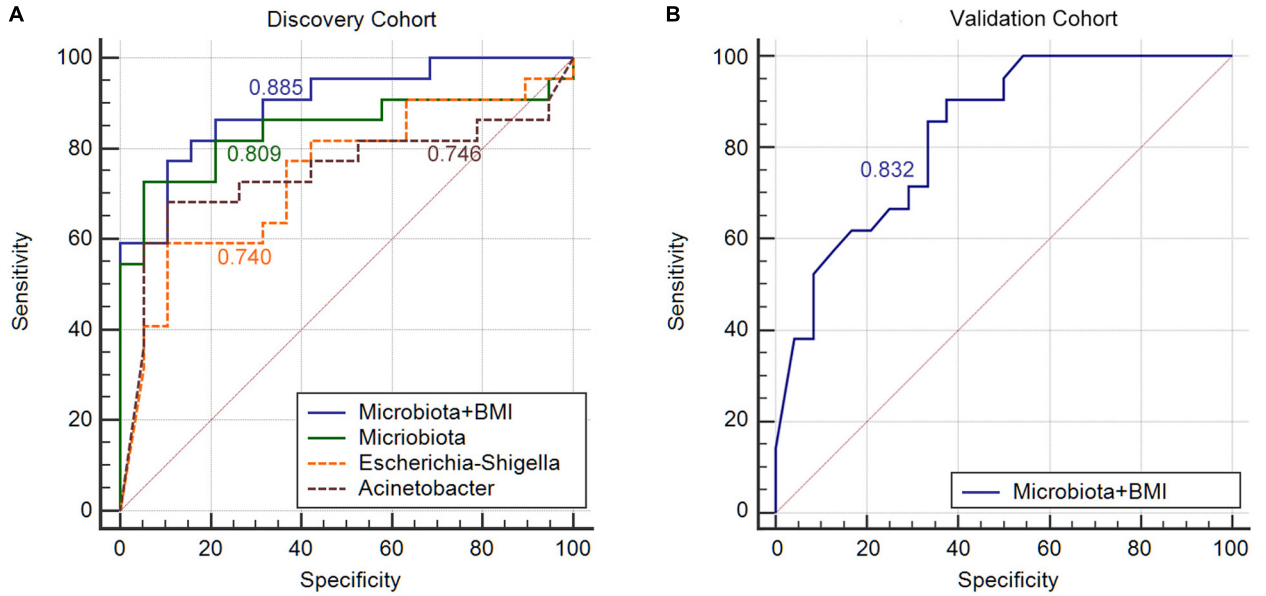


FIGURE 4 | Gut microbiota signature can be used to discriminate between MA patients from nMA patients. **(A)** ROC analysis in discovery cohort with *Escherichia–Shigella* along, *Acinetobacter* along, combination of two genera (Microbiota), and bacterial genera together with BMI (Microbiota + BMI). **(B)** ROC analysis with bacterial genera together with BMI (Microbiota + BMI) in validation cohort.

TABLE 4 | Sensitivity, specificity, PPV and NPV of the risk score based on predominant presence of the risk factors.

Risk score	Discovery cohort					Validation cohort				
	Sensitivity (%)	Specificity (%)	PPV	NPV	MA rate*	Sensitivity (%)	Specificity (%)	PPV	NPV	MA rate*
0	100	0	53.7	/	0 (0/6)	100	0	46.67	/	0 (0/2)
1	100	31.6	62.9	100	15.38 (2/13)	100	8.33	48.84	100	14.28 (1/7)
2	90.9	89.5	90.9	89.5	83.3 (10/12)	95.24	33.33	55.56	88.89	23.08 (3/13)
3	45.5	100	100	61.3	100 (5/5)	80.95	75	73.91	81.82	66.67 (10/15)
4	22.7	100	100	52.8	100 (5/5)	33.33	95.83	87.5	62.16	87.5 (7/8)

samples from resected tumor after operation may be disturbed by clinical intervention, such as the preventive antibiotics application before operation. A clear clustering between the MA and nMA patients was observed. α-diversity of the mucosal and fecal samples were both lower in the MA group. As low diversity microbiota indicated unstable ecosystem, one piece of evidence that has emerged from many large surveys of gut microbial communities is that low microbial diversity is almost invariably associated with disease (Round and Palm, 2018).

It was noticed that there were predominated bacterial taxa in both MA and nMA, respectively. Specifically, we found the genera enriched in nMA group positively correlated each other. This co-abundance groups (CAG) of bacterial taxa resembled the previously formulated concept of enterotypes. The bacterial taxa belonged to one CAG may relate to each other not only quantitatively but also functionally (Flemer et al., 2017). *Escherichia–Shigella* was identified as the most abundant genus in the MA patients. *Escherichia* comprises eight species, including the well-known *Escherichia coli* (*E. coli*). Although *Shigella* is technically a independent genus with four species, they are inseparable from *E. coli* in terms of 16S rRNA gene DNA sequence, so they are commonly bracketed together and named *Escherichia–Shigella* in 16S rRNA gene-based microbiota studies. All these species belong to the Enterobacteriaceae, which was highly enriched in the MA patients as well. *Escherichia–Shigella* has been shown to produce Colibactin, which is encoded by polyketide synthase (*pks*) genotoxicity island (Nougayrède et al., 2006). Colibactin possesses the capacity to damage DNA and lead to CRC development (Wu et al., 2009; Arthur et al., 2012). Mucosa-associated *E. coli* has been found to be significantly more prevalent in CRC tissue and correlates with tumor stage and prognosis (Bonnet et al., 2014).

E. coli and *Shigella* have been shown to increase intestinal permeability in this intestinal disorder, likely due to down-regulation of tight junction proteins (Cinova et al., 2011). Our study demonstrated that *Escherichia–Shigella* was positively correlated with *bacterial invasion of epithelial cells pathway*, which was also enriched in the MA patients as identified by PICRUSt method. The *bacterial invasion of epithelial cells pathway* indicates that

the potential pathogens such as *Escherichia–Shigella* and *Enterococcus* could adhere the surface of host cells, cross host epithelial barriers, and get access to internal tissues, thereby promoting their dissemination inside the host (Ribet and Cossart, 2015).

It was striking that there was high similarity in the mucosal microbiota of paired on–off tumor samples with regard to overall composition of the microbiota. In contrast, paired fecal and mucosal samples had lower similarity. These findings were consistent with a previous study (Nakatsu et al., 2015). We found that microbiota in the fecal samples can be also separated between the MA and nMA groups. As such, even though fecal microbiota differed from and may only partially reflects the microbiota at the mucus layer, differences due to MA status are still evident. Unlike mucosal samples, which mainly reflected the local microbiota, the fecal samples may be a representative for the whole gut environment. It is possible that except for the lesion site, other sites of the colon may also possess more CRC-related bacteria in the MA patients, compared to the nMA patients.

Our previous cross-sectional study showed differences in post-operative fecal microbiota between patients with and without MA (Jin et al., 2019). We wonder whether such difference could exist in the pre-operative fecal samples. As observed in our study, similar to pre-operative CRC cohort, lower microbiota diversity, and higher abundance CRC-related bacterial taxa were characteristics for MA in the post-operative cohort, but not obvious as pre-operative cohort. ANOSIM results also showed the distance value between MA and nMA was high in pre-operative cohort. Collectively, these findings suggest residual microbiota features for MA still exist in post-operative cohort.

In this study, we identified novel microbiome biomarkers for prediction of the MA. It is important to highlight that MA is a complex disease that occurs as a combination of microbial colonization, patient genetic background, and other environment factors. Given that, we established the RF model utilizing the gut microbiota together with the clinical risk factors to predict MA. We observed that the key predictor was *Escherichia–Shigella* in this model which was in agreement with logistic regression result, showing that *Escherichia–Shigella* was an independent risk factors with an overt OR value of 53.254. Although synchronous adenoma was not included in the RF model, in view of it as a risk factor for MA and in order to translate our result to clinical application, we developed a risk score based on presence of the negative prognostic genus *Escherichia–Shigella*, absence of the positive prognostic genus *Acinetobacter*, together with high BMI and the traditionally accepted risk factors, synchronous adenoma. The specificity was lower in the validation cohort; one explanation maybe the discovery cohort derived cut-off value was not optimized enough, but there was still a high sensitivity in validation cohort and the overall AUC value was reasonable. As expected, the RF model performed well for off-tumor mucosal and fecal samples. The RF model cannot predict local recurrence with data imported from Bullman et al. (2017) study, which may indicate the specificity of our predict model. Although this clinical condition is an excellent model for investigating whether dysbiosis precedes MA, we can't draw conclusions regarding the causality on the basis of our data. We wonder if CRC patients at high risk for MA could be identified pre-operatively by gut microbiota; an individual post-operative surveillance plan can be made to prevent the occurrence of metachronous CRC.

Our study may have a number of limitations. Firstly, patients were followed up, but mucosal or fecal samples were not collected after surgery, for which we cannot make a before–after analysis in the same cohort of patients. But, we made conjoint analysis with previous data of another cohort patients. Secondly, the sample size was relatively small, and the predicted potential of the selected biomarkers should be evaluated in an independent cohort. Although no external cross-validation was achieved in this study, sufficient internal cross-validation with different samples was made. Thirdly, the patients were followed up with for 12 months, so we could only observe MA development, but not metachronous carcinoma.

The findings have demonstrated that specific members of the dominant gut microbiota as non-invasive biomarkers for prediction of MA or CRC after surgical resection. The newly established RF algorithm and the risk-scoring system have a good ability to predict the development of MA after surgical resection, and therefore, the novel approaches hold potential to guide individual post-operative surveillance plan for CRC patients in future clinical application.

DATA AVAILABILITY STATEMENT

The raw sequences have been deposited in the NCBI Sequence Read Archive (Nos. PRJNA594545 and PRJNA573487), and the necessary metadata can be found at https://www.ncbi.nlm.nih.gov/Traces/study/ by searching the respective SRA study accession.

ETHICS STATEMENT

The study protocol was reviewed and approved by the Research Ethics Committee of the First Affiliated Hospital of Harbin Medical University. Each patient had provided a written informed consent. The study involving human subjects was strictly performed according to international guidelines regarding the conduct of clinical trials. This study was registered at ClinicalTrials.gov (NCT03667495).

AUTHOR CONTRIBUTIONS

YL and YW conceived the study design. RG and LL recruited and followed up the patients. XJ coordinated with patients transported patient samples. YL and FZ performed the sequencing analysis. WY and SW contributed to the data analyses. XG and GG maintained patient records. YL and YW drafted the manuscript. All authors read and approved its final version.

FUNDING

This work was supported by grant 81970466 from National Natural Science Foundation of China.

ACKNOWLEDGMENTS

We thank the patients enrolled in the study for their commitment with the project. Sequence processing and analysis were performed using the resources of the Majorbio Bio-Pharm Technology (Shanghai, China). This manuscript has been released as a pre-print at [ResearchSquare] (Yang et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01106/full#supplementary-material

REFERENCES

Arthur, J. C., Perez-Chanona, E., Muhlbauer, M., Tomkovich, S., Uronis, J. M., Fan, T. J., et al. (2012). Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* 338, 120–123.

Balleste, B., Bessa, X., Pinol, V., Castellvi-Bel, S., Castells, A., Alenda, C., et al. (2007). Detection of metachronous neoplasms in colorectal cancer patients: identification of risk factors. *Dis. Colon. Rectum.* 50, 971–980. doi: 10.1007/s10350-007-0237-2

Bonnet, M., Buc, E., Sauvanet, P., Darcha, C., Dubois, D., Pereira, B., et al. (2014). Colonization of the human gut by *E. coli* and colorectal cancer risk. *Clin. Cancer Res.* 20, 859–867. doi: 10.1158/1078-0432.ccr-13-1343

Bouvier, A. M., Latournerie, M., Jooste, V., Lepage, C., Cottet, V., and Faivre, J. (2008). The lifelong risk of metachronous colorectal cancer justifies long-term colonoscopic follow-up. *Eur. J. Cancer* 44, 522–527. doi: 10.1016/j.ejca.2008.01.007

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., et al. (2017). Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science* 358, 1443–1448.

Buttigieg, P. L., and Ramette, A. (2014). A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol. Ecol.* 90, 543–550. doi: 10.1111/1574-6941.12437

Cinova, J., De Palma, G., Stepankova, R., Kofronova, O., Kverka, M., Sanz, Y., et al. (2011). Role of intestinal bacteria in gliadin-induced changes in intestinal mucosa: study in germ-free rats. *PLoS One* 6:e16169. doi: 10.1371/journal.pone.0016169

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* 6:6528.

Flemer, B., Lynch, D. B., Brown, J. M., Jeffery, I. B., Ryan, F. J., Claesson, M. J., et al. (2017). Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* 66, 633–643. doi: 10.1136/gutjnl-2015-309595

Jin, Y., Liu, Y., Zhao, L., Zhao, F., Feng, J., Li, S., et al. (2019). Gut microbiota in patients after surgical treatment for colorectal cancer. *Environ. Microbiol.* 21, 772–783.

Kahi, C. J., Boland, C. R., Dominitz, J. A., Giardiello, F. M., Johnson, D. A., Kaltenbach, T., et al. (2016). Colonoscopy surveillance after colorectal cancer resection: recommendations of the US multi-society task force on colorectal cancer. *Gastroenterology* 150, 758.e11–768.e11.

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., et al. (2013). *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 14, 207–215. doi: 10.1016/j.chom.2013.07.007

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Meyerhardt, J. A., Mangu, P. B., Flynn, P. J., Korde, L., Loprinzi, C. L., Minsky, B. D., et al. (2013). Follow-up care, surveillance protocol, and secondary prevention measures for survivors of colorectal cancer: american Society of Clinical Oncology clinical practice guideline endorsement. *J. Clin. Oncol.* 31, 4465–4470. doi: 10.1200/jco.2013.50.7442

Mulder, S. A., Kranse, R., Damhuis, R. A., Ouwendijk, R. J., Kuipers, E. J., and van Leerdam, M. E. (2012). The incidence and risk factors of metachronous colorectal cancer: an indication for follow-up. *Dis. Colon. Rectum.* 55, 522–531. doi: 10.1097/dcr.0b013e318249db00

Nakatsu, G., Li, X., Zhou, H., Sheng, J., Wong, S. H., Wu, W. K., et al. (2015). Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat. Commun.* 6:8727.

Nougayrède, J. P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., et al. (2006). *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* 313, 848–851. doi: 10.1126/science.1127059

Ribet, D., and Cossart, P. (2015). How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes Infect.* 17, 173–183. doi: 10.1016/j.micinf.2015.01.004

Round, J. L., and Palm, N. W. (2018). Causal effects of the microbiota on immune-mediated diseases. *Sci. Immunol.* 3:eaao1603. doi: 10.1126/sciimmunol.aao1603

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60.

Sze, M. A., Baxter, N. T., Ruffin, M. T. T., Rogers, M. A. M., and Schloss, P. D. (2017). Normalization of the microbiota in patients after treatment for colonic lesions. *Microbiome* 5:150.

van der Stok, E. P., Spaander, M. C. W., Grünhagen, D. J., Verhoef, C., and Kuipers, E. J. (2016). Surveillance after curative treatment for colorectal cancer. *Nat. Rev. Clin. Oncol.* 14, 297–315.

Wu, S., Rhee, K. J., Albesiano, E., Rabizadeh, S., Wu, X., Yen, H. R., et al. (2009). A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat. Med.* 15, 1016–1022. doi: 10.1038/nm.2015

Yang, L., Rui, G., Lujia, L., Xiangren, J., Wei, Y., Fuya, Z., et al. (2019). *GutMicrobiota-Based Algorithms in the Prediction of Metachronous Adenoma in Colorectal Cancer Patients Following Surgery*. Available online at: https://www.researchsquare.com/article/rs-8196/v1 (accessed November 1, 2019).

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 November 2020
doi: 10.3389/fmicb.2020.583525

A Pilot Study: Changes of Intestinal Microbiota of Patients With Non-small Cell Lung Cancer in Response to Osimertinib Therapy

Jing Cong[1], Yuguang Zhang[2], Yadong Xue[2], Chuantao Zhang[3], Mingjin Xu[3], Dong Liu[3], Ruiyan Zhang[4] and Hua Zhu[3]*

[1] College of Marine Science and Biological Engineering, Qingdao University of Science and Technology, Qingdao, China, [2] Key Laboratory of Forest Ecology, Environment of State Forestry Administration, Institute of Forestry Ecology, Environment and Protection, Chinese Academy of Forestry, Beijing, China, [3] Department of Oncology, The Affiliated Hospital of Qingdao University, Qingdao University, Qingdao, China, [4] Department of Radiotherapy, Qingdao Central Hospital, Qingdao, China

OPEN ACCESS

Edited by:
Steve Lindemann, Purdue University, United States

Reviewed by:
Pallavi Singh, Northern Illinois University, United States
Tarique Hussain, Nuclear Institute for Agriculture and Biology, Pakistan

***Correspondence:**
Jing Cong
yqdh77@163.com

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *15 July 2020*
Accepted: *26 October 2020*
Published: *10 November 2020*

Citation:
Cong J, Zhang Y, Xue Y, Zhang C, Xu M, Liu D, Zhang R and Zhu H (2020) A Pilot Study: Changes of Intestinal Microbiota of Patients With Non-small Cell Lung Cancer in Response to Osimertinib Therapy. Front. Microbiol. 11:583525. doi: 10.3389/fmicb.2020.583525

Osimertinib contributes to the higher efficacy and few intestinal side effects in non-small cell lung cancer (NSCLC) patients with T790M mutation. Previous studies has reported that intestinal microbiota play important roles in drug efficacy and toxicity. However, we have known less about the changes of intestinal microbiota in response to osimertinib therapy. In this pilot study, we used longitudinal sampling with 6 weeks sampling collection intervals for about 1 year to model intestinal microbial changes based on the 16S rRNA genes sequencing in fecal samples from NSCLC patients in response to osimertinib therapy. The results showed that there was no significantly different on the intestinal microbial composition at the phylum, family, and genus level among NSCLC patients with different treatment cycles ($P > 0.05$). There were no significant differences in alpha diversity characterized by the richness, Shannon diversity, and phylogenetic diversity based on the Welch's *t*-test among NSCLC patients in response to osimertinib therapy ($P > 0.05$). However, the dissimilarity test and principal coordination analysis showed a few differences among NSCLC patients. The intestinal microbial markers were changed in post-therapy (*Sutterella*, *Peptoniphilus*, and *Anaeroglobus*) compared to that in pre-therapy (*Clostridium XlVa*). Furthermore, the phylogenetic molecular ecological networks (MENs) were influenced by osimertinib therapy based on the module number, link number, and module taxa composition of the first six groups. Overall, it indicated that osimertinib therapy changed the intestinal microbiota to some extent, though not completely. In all, this pilot study provides an understanding of changes of intestinal microbiota from NSCLC patients in response to osimertinib therapy. No complete changes in intestinal microbiota seem to be closely linked with the few intestinal side effects and higher efficacy in response to osimertinib therapy.

Keywords: intestinal microbiota, 16S rRNA sequencing, non-small cell lung cancer patients, osimertinib therapy, ecological network analysis

INTRODUCTION

Lung cancer remains the leading cause of cancer-related deaths worldwide (Torre et al., 2016). Non-small cell lung cancer (NSCLC) accounts for most of all cases of lung cancer, including adenocarcinoma, squamous cell carcinoma, and large-cell lung cancer, which is generally diagnosed at a terminal stage of lung cancer. For a long time, platinum-based chemotherapy has represented the cornerstone for the first-line treatment of advanced NSCLC patients (Santarpia et al., 2017a), although with several limitations, including a number of side effects and a dismal overall survival. In recent years, the development of specific molecularly targeted agents has primarily changed the therapeutic landscape for advanced NSCLC patients, including epidermal growth factor receptor-tyrosine kinase inhibitors (EGFR-TKIs)-, anaplastic lymphoma kinase (ALK)-, and BRAF-inhibitors (Rosell and Karachaliou, 2016). These therapies have greatly improved the survival and quality of NSCLC patients. Gefitinib, afatinib, and erlotinib are the standard first-line treatment for advanced EGFR mutated NSCLC patients. After a variable length of time from starting treatment, the resistance mechanisms of first- and second- generation EGFR-TKIs inevitably emerge. The T790M mutation at exon 20 within the kinase domain of EGFR is the most common mechanism of acquired resistance, which occurs in approximately 50–60% of EGFR-TKI-resistant tumors.

Osimertinib is the third-generation for the treatment of patients with metastatic EGFR T790M-positive NSCLC (Cross et al., 2014), which is the first compound granted US Food and Drug Administration (FDA) and European Medicine Agency (EMA) approval (Santarpia et al., 2017b). Soria et al. (2018) found that the advanced NSCLC patients with previously untreated, EGFR mutation-positive receiving osimertinib had the significantly longer median progression-free survival (PFS) than those receiving gefitinib or erlotinib in a double-blind, phase 3 trial (18.9 vs. 10.2 months; $P < 0.001$). The median overall survival (OS) was 38.6 months in response to osimertinib therapy and 31.8 months in response to gefitinib or erlotinib therapy (Ramalingam et al., 2020). Furthermore, there were less adverse events of grade 3 or higher in the osimertinib group than that in the comparator group (34 vs. 45%) (Soria et al., 2018). Mok et al. (2017) reported that the median duration of PFS in these T790M-positive advanced NSCLC patients with osimertinib, who had disease progression after first-line EGFR-TKI therapy, was significantly longer than those with platinum therapy plus pemetrexed (10.1 vs. 4.4 months; $P < 0.001$). The less adverse events of grade 3 or higher were lower with osimertinib compared to the platinum therapy plus pemetrexed (23 vs. 47%) (Mok et al., 2017). Based on its significant efficacy, safety and favorable toxicity profile, osimertinib has been considered as a therapeutic option preferable to early generation EGFR-TKI for further improving the clinical outcome of EGFR-mutated patients (Mok et al., 2017). However, the disease would progress after receiving osimertinib therapy for approximately 10 months. Thus, some novel therapeutic strategies should overcome the osimertinib resistance.

Recently, intestinal microbiota has emerged as an "organ" that plays a key role in health and disease. The intestinal microbial composition shows high inter-individual variations (Huttenhower et al., 2012). Various studies have proved that effect of drug intake in intestinal microbiota (Wu et al., 2017). In turn, intestinal microbiota can also contribute to the different in response to a specific drug in different individuals (Wu et al., 2017). The intestinal microbiota can directly transform the drug or change the host's metabolism and immune system to modify the pharmacodynamics of a medication (Rajpoot et al., 2018). Therefore, understanding the role of intestinal microbiota in drug response may contribute to the development of microbiome-targeting approaches that improve the drug efficacy.

Previous studies has reported that the intestinal microbiota play important roles in drug efficacy and toxicity in response to chemotherapeutic drugs (Alexander et al., 2017; Cong et al., 2019). These studies showed that drugs could change the composition of intestinal microbiota for patients. However, these works drew the conclusion with only a few times points. The dynamic patterns of microbial communities across longer time scales with drug usage remain unclear. In this study, we longitudinally tracked the changes of intestinal microbiota in NSCLC patients in response to targeted drug osimertinib therapy for nine cycles based on the 16S rRNA sequencing data. This pilot study will help the development of personalized medicine, and try to modulate the intestinal microbiota to manage drug efficiency on the level of the individual (Doestzada et al., 2018).

MATERIALS AND METHODS

Study Subjects and Microbial Sampling

Non-small Cell Lung Cancer (NSCLC) Patients

Eight adults with locally advanced (stage IIIB) or metastatic (stage IV) NSCLC with confirmed T790M mutation from the Affiliated Hospital of Qingdao University, who have received prior EGFR-TKI therapy, were recruited to the pilot study (**Figure 1**). Exclusion criteria included that those who have inflammatory bowel disease, irritable bowel syndrome, and other intestinal diseases, and were treated with antibiotics or probiotics usage, during the sampling. Sampling time was ranged from April 2017 to May 2018. Daily dosage of osimertinib was administered orally as one 80 mg tablet. Total of 65 fecal samples were collected from nine cycles of osimertinib therapy. The sample collection intervals of every cycle were about 6 weeks. Fecal samples were self-sampled in the morning prior to the start of drug usage, including that before the first therapy, before the second therapy, before the third therapy, before the fourth therapy, before the fifth therapy, before the sixth therapy, before the seventh therapy, before the eighth therapy, before the ninth therapy, and before the tenth therapy, named by T1, T2, T3, T4, T5, T6, T7, T8, T9, and T10, respectively.

Healthy Individuals

Control samples were obtained from 21 healthy individuals. These healthy individuals, who have any recorded antibiotics or probiotics usage, and gastrointestinal tract disorders within

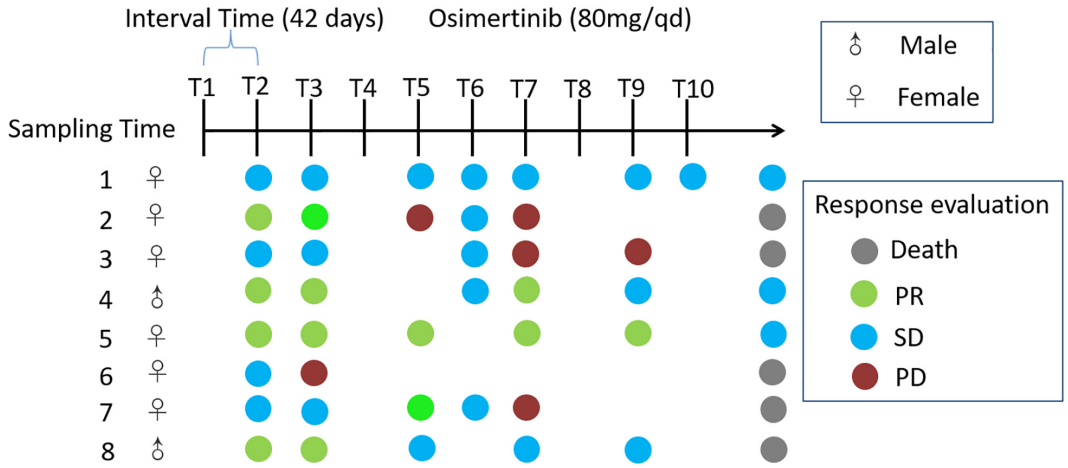


FIGURE 1 | Timeline representing the clinical course of eight patients. Key time points include timing of fecal sampling; sampling time was ranged from April 2017 to May 2018; Response evaluations include partial response (PR), stable disease (SD), progression disease (PD) and death. Response evaluations were due to December 2018. Daily dosage of osimertinib was administered orally as one 80 mg tablet.

1 month preceding the sample collection, were excluded. The collected samples from the healthy individuals were named by H.

All of the study subjects have been local residents of Qingdao city. This pilot study was approved by the Affiliated Hospital of Qingdao University Institutional Review Board, and all pilot study subjects signed the informed consent before participation. Fresh fecal samples were put into 5 ml tubes and immediately stored at −80°C until the day of analysis.

DNA Extraction, PCR Amplification of 16S rRNA Gene, Amplicon Sequencing and Data Processing

Total genomic DNA was extracted using the DNA Stool Kit from Tiangen (Wu et al., 2016) and purity were monitored on 1% agarose gels. 16S rRNA gene of V3–V4 regions amplicon sequencing was carried out employing the 16S Metagenomic Sequencing Library Preparation protocol developed by Illumina (San Diego, California, United States) using the bacterial universal primers (357F-806R) (Cong et al., 2018). The PCR amplification products were purified with Qiagen Gel Extraction Kit (Qiagen, Germany). The DNA quality was assessed on the Qubit® 2.0 Fluorometer (Thermo Fisher Scientific) and Agilent Bioanalyzer 2100 system. Finally, bacterial DNA amplicons were sequenced from each fecal sample for 2 × 250 bp paired-end sequencing based on the Illumina Hiseq 2500.

16S rRNA Amplicon Sequencing Data Analysis

Raw sequences were separated into samples by barcodes based on the Galaxy Illumina sequencing pipeline. Ambiguous, adapters, and low-quality reads ("N") were trimmed by Btrim (Kong, 2011). Forward and reverse reads were incorporated into a whole sequence by FLASH (Magoc and Salzberg, 2011). After quality control of the raw data, the clean reads were clustered into operational taxonomic units (OTUs) by using UCLUST at 97% similarity level (Edgar, 2010). Each OTU was considered to represent a species (Deng et al., 2012). The ribosomal database project (RDP) classifier was used to determine the taxonomic assignment (Wang et al., 2007). Rarefaction analysis was performed using the original detected OTUs.

Network Analysis

Intestinal microbial ecological networks were constructed and analyzed by random matrix theory (RMT) methods based on the online MENA pipeline. OTUs detected in more than half in each group were used to ensure reliable correlations. To compare with different networks, the same cutoff of 0.85 was applied to construct ecological networks for intestinal microbial communities. Each network was divided into modules by the fast greedy modularity optimization to describe the modularity property. In addition, a network developed by OTU abundance data represented the ecological links of different OTU nodes (OTUs) in a microbial community, and different nodes played distinct roles (Guimerà et al., 2007).

Statistical Analysis

The Shannon index, richness, and phylogenetic diversity were calculated for alpha diversity analysis, which presented complexity of species diversity for samples. The different tests of alpha diversity for different groups were performed by Wilcoxon Rank Sum Test. Beta diversity was calculated by Bray-Curtis distance. Differences in beta diversity were identified using the multiple response permutation procedure (MRPP) algorithm. Community structure based on beta diversity was visualized using principal coordinate analysis (PCoA). Linear Discriminant Analysis with Effect Size (LEfSe) was used to identify the significant *P*-values associated with microbial clades and functions. Characteristics with a LDA score cutoff of 2.0 were known as being different. Significantly different biomarkers at the phylum and genus levels were identified using STAMP (v2.1.3).

An absolute Pearson's correlation was based on a significance level under 0.05. Principal components analysis (PCA) was used to determine the changes of intestinal microbiota based on significant different genera and OTUs. The R software package (v3.4.1) was used for all statistical analysis, except for two-tailed unpaired *t*-tests and Pearson correlation by IBM SPSS statistic 19.0 to determine the significance of the differences.

RESULTS

Study General Characteristic

NSCLC patients and control subjects were matched for age, sex as well as body mass index (BMI) in this study (**Table 1**). Eight NSCLC patients with locally advanced (stage IIIB) or metastatic (stage IV) NSCLC treated with osimertinib therapy after progression were enrolled in the present study. Response evaluation was administered every 6 weeks. During the treatment, no antibiotics were applied. Patients were classified based on radiological evaluation according to Response Evaluation Criteria in Solid Tumors (RECIST 1.1) (Schwartz et al., 2016). Fecal samples were collected every 6 weeks during therapy until disease progression, or death, or the study self-withdrawal, according to the informed consent and study protocol. Dynamic variation of intestinal bacterial characteristics was evaluated and analyzed by metagenomic sequencing.

Changes of Intestinal Microbial Composition at the Taxonomical Level From NSCLC Patients in Response to Osimertinib Therapy

A total of 678 OTUs were defined with RDP annotations, including 351 OTUs belonging to 109 genera, and 327 OTUs of unclassified genera. Rarefaction curves showed that most samples leveled out between 100 and 250 taxa (**Supplementary Figure S1**). At the phylum level, the distribution pattern of the top six phylotypes (comprising about 99% of the total counts) in each group is shown in **Supplementary Figure S2**. We explored the differences of phylum Bacteroidetes, Firmicutes, Proteobacteria, Verrucomicrobia, Actinobacteria, and Fusobacteria in NSCLC patients and healthy individuals (**Supplementary Table S1**). The NSCLC samples showed no obvious differences in relative abundance of these phylotypes between pre-therapy and post-therapy ($P > 0.05$). There were also no significant different between NSCLC samples and healthy samples ($P > 0.05$). At the family level, Bacteroidaceae, Lachnospiraceae, and Prevotellaceae were the top three family almost in NSCLC patients and healthy individuals (**Supplementary Figure S3**). There were almost no significant differences in relative abundance of selected taxa between pre-therapy samples and post-therapy samples, and between NSCLC samples and healthy samples ($P > 0.05$, **Supplementary Table S2**). At the genus level, *Bacteroides* was the most abundant genus, followed by *Prevotella* in both NSCLC patients and healthy individuals, except for T9 and T10 (**Supplementary Figure S4**). We also explored the differences of the genus *Bacteroides*, *Prevotella, Faecalibacterium*, and other sixteen genera in NSCLC patients in response to osimertinib therapy. The results also showed that no significant differences were detected between pre- and post-therapy samples, and between NSCLC patients and healthy individuals ($P > 0.05$, **Supplementary Table S3**).

Changes of Intestinal Microbial Diversity From NSCLC Patients in Response to Osimertinib Therapy

As measures of alpha diversity (**Supplementary Figure S5**), which describes diversity within each sample, we used richness (number of distinct species present in samples), phylogenetic diversity, and Shannon diversity to explore the changes in eight patients in response to osimertinib therapy. The results showed that alpha diversity of individuals changed greatly (**Supplementary Figure S6**). Most samples in group T showed less species richness, phylogenetic diversity, and Shannon diversity than those in group H (**Table 2**). The Welch's *t*-test showed almost no significant differences between group H and group T, and between pre-therapy and post-therapy ($P > 0.05$, **Supplementary Table S4**). Dissimilarity analysis showed significant differences between pre-therapy (T1) and post-therapy (T2, T3, T4, T5, T6, T7, T8, T9, T10) based on the MRPP ($P < 0.05$, **Supplementary Table S5**). However, no significant differences in group H and group T ($P > 0.05$, **Supplementary Table S5**). Principal coordination analysis based on Bray-curtis dissimilarity index showed a little separation between healthy individuals and NSCLC patients (**Figure 2**).

Differences of Intestinal Microbiota From NSCLC Patients in Response to Osimertinib Therapy

To identify intestinal microbial responses associated with osimertinib therapy at the taxonomical level, we determined

TABLE 1 | Characteristics of the study subjects and the samples.

Subject group	No. of subjects (male/female)	Mean age (range)	Mean BMI (range)	Treatment period (sample number)									
				T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Healthy individuals (H)	21 (4/17)	54 (26–64)	23.4	21									
Non-small cell lung cancer patients (T)	8 (2/6)	60 (52–68)	24.4	8	8	8	7	7	7	6	5	5	4

microbial clade differences using LEfSe analysis (**Figure 3**). At the phylum level, we found that the higher proportions of Actinobacteria were observed in H than that in T1 and T10 (**Supplementary Figure S7**). At the genus level, greater proportions of *Bacteroides*, *Klebsiella*, and *Parasutterella* were detected in H than that in T1 and T10 (**Supplementary Figure S7**). The genera *Clostridium XIVa* and *Cellulosilyticum* were significantly enriched in T1 than that in H and T10 (**Figure 3**). The members of *Sutterella*, *Peptoniphilus*, *Anaeroglobus*, and *Neisseria* were more abundant in T10 than that in T1 and H (**Figure 3**). In addition, we constructed the PCA plot based on the significant different genera and OTUs in group T and group H (**Figure 4**). The results showed that the samples from NSCLC patients were well separated from the healthy individuals, but partly overlapped within different treatment cycles based on the different genera and OTUs (**Figures 4A,B**). We also selected the group T1, T3, T5, T7, and T10 to structure the PCA plot based on the significant distinct genera and OTUs (**Supplementary Figure S8**). It also showed the changes in response to osimertinib therapy.

Molecular Ecological Network Analysis of Intestinal Microbiota From NSCLC Patients in Response to Osimertinib Therapy

The molecular ecological networks (MENs) were constructed for NSCLC patients to determine the effect of osimertinib therapy on microbial assemblages that potential interact with intestinal niches. We focused on representative networks from NSCLC patients with more than six biological duplications, including of the group T1, T2, T3, T4, T5, and T6. No less than five nodes to construct the modules in NSCLC samples (**Figure 5**). There were 3, 1, 1, 3, 1, and 1 module(s) in group T1, T2, T3, T4, T5, and T6 networks, respectively (**Supplementary Table S6**). Overall, taxa tended to co-occur (positive correlations, pink lines) rather than co-exclude (negative correlations, blue lines) (**Figure 5**). The negative correlations accounted for less than 45% of the potential interactions observed at each treatment stage (**Figure 5**). The negative correlations in NSCLC patients were increased by 22.32% from T1 to T6. The composition of the modules differed within each network and changed over the treatment time (**Figure 5**). Firmicutes almost dominated all the modules from each treatment stage in NSCLC patients. The phylum Fusobacteria presented in the modules before the third treatment (T3) and before the sixth treatment (T6). The phylum Fusobacteria was supposed to be more relevant to intestinal dysbiosis (Brennan and Garrett, 2019).

TABLE 2 | The richness, phylogenetic diversity, and Shannon diversity in non-small cell lung cancer patients and healthy individuals.

Group	Richness	Phylogenetic diversity	Shannon diversity
H	161 ± 39	12.25 ± 2.18	4.44 ± 0.83
T1	125 ± 39	9.84 ± 2.65	3.97 ± 0.96
T2	136 ± 56	10.87 ± 3.34	3.63 ± 1.82
T3	127 ± 55	9.81 ± 3.48	4.20 ± 0.85
T4	142 ± 47	11.01 ± 3.45	4.45 ± 0.83
T5	153 ± 41	11.52 ± 2.49	4.72 ± 0.53
T6	148 ± 61	11.61 ± 3.63	4.30 ± 1.34
T7	127 ± 49	10.14 ± 3.32	4.10 ± 0.82
T8	132 ± 46	10.16 ± 2.73	4.15 ± 0.69
T9	123 ± 41	10.03 ± 2.42	3.78 ± 0.78
T10	127 ± 64	10.12 ± 4.19	3.67 ± 1.33

H represents the healthy individuals; Tn (n = 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10) represents the non-small cell lung cancer patients in different cycles of osimertinib therapy.

DISCUSSION

The identification of tyrosine kinase inhibitors (TKIs), has marked the advent of the era of precision medicine, which has revolutionized the diagnostic and therapeutic approach to NSCLC. Recently, osimertinib, which is designed to preferentially target sensitizing mutations and the T790M resistance mutation, over the wild-type receptor, has significantly improved survival and quality of life in molecularly defined subgroups of NSCLC patients. It has been reported that intestinal microbiome and drugs or drug metabolites interact with intestinal and systemic pharmacological effects. Intestinal microbiota play key roles in compound modifications including their activation (Tiago et al., 2014), inactivation (Haiser et al., 2013), or toxification (Wallace et al., 2010; Zimmermann and Zimmermann-Kogadeeva, 2019). In turn, the drug metabolite could change composition and structure. In this study, we explored that the changes of intestinal microbiota with NSCLC patients in response to osimertinib therapy for nine cycles.

Firstly, we examined the changes of intestinal microbial composition from NSCLC patients at the phylum and genus level in response to osimertinib therapy. Generally, the gut was dominated by members of four bacterial phyla, Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria, with lesser and sporadic representation of other phyla, such as Fusobacteria and Verrucomicrobia (Rajilic-Stojanovic and De Vos, 2014; Budden et al., 2017; Wexler and Goodman, 2017). Consistent with it, our results demonstrated that the majority of all microbial populations identified in our participants were Bacteroidetes, Firmicutes, and Proteobacteria (**Supplementary Figure S2**). Previous numerous studies have characterized the lung microbiome using bronchoalveolar lavage microbiota of subjects with lung diseases. Significant differences are found in bacterial community composition between healthy and diseased lungs (Garzoni et al., 2013; Dickson et al., 2014). An increasing number of studies have revealed the close relationship between the intestinal microbial composition and lung diseases, known as the gut-lung axis (Budden et al., 2017; Zhuang et al., 2019). For example, an increase in the abundance of *Bacteroides fragilis* and total anaerobes, as well as a decrease in the relative abundance of *Faecalibacterium* spp., *Veillonella* spp., *Rothia* spp., and *Lachnospira* spp. in early life were associated with increased risk of asthma (Vael et al., 2008; Arrieta et al., 2015). Recently, Strickertsson et al. (2013), Amarnani and Rapose

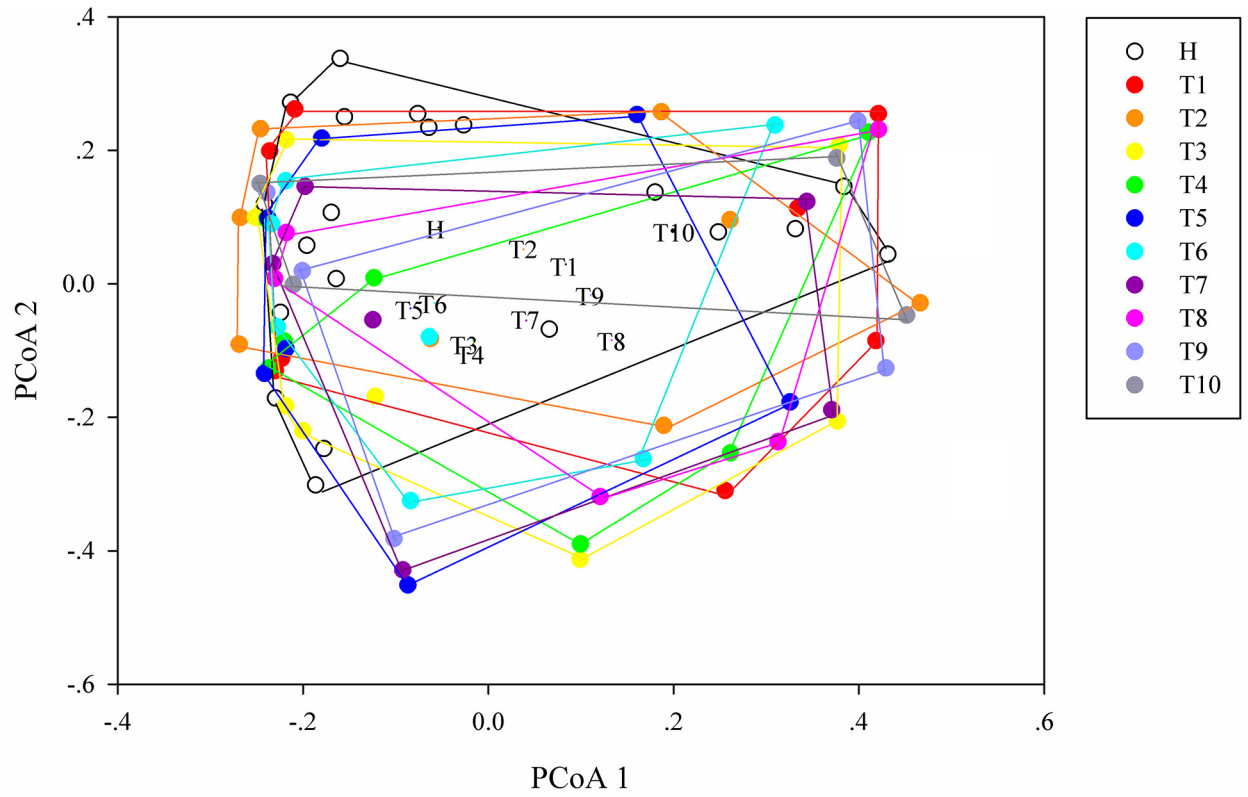


FIGURE 2 | Principal coordinates analysis (PCoA) ordination (operational taxonomic units = 97% 16S rRNA sequence similarity) showing different microbial composition between NSCLC patients and healthy individuals based on the Bray-Curtis dissimilarity matrix.

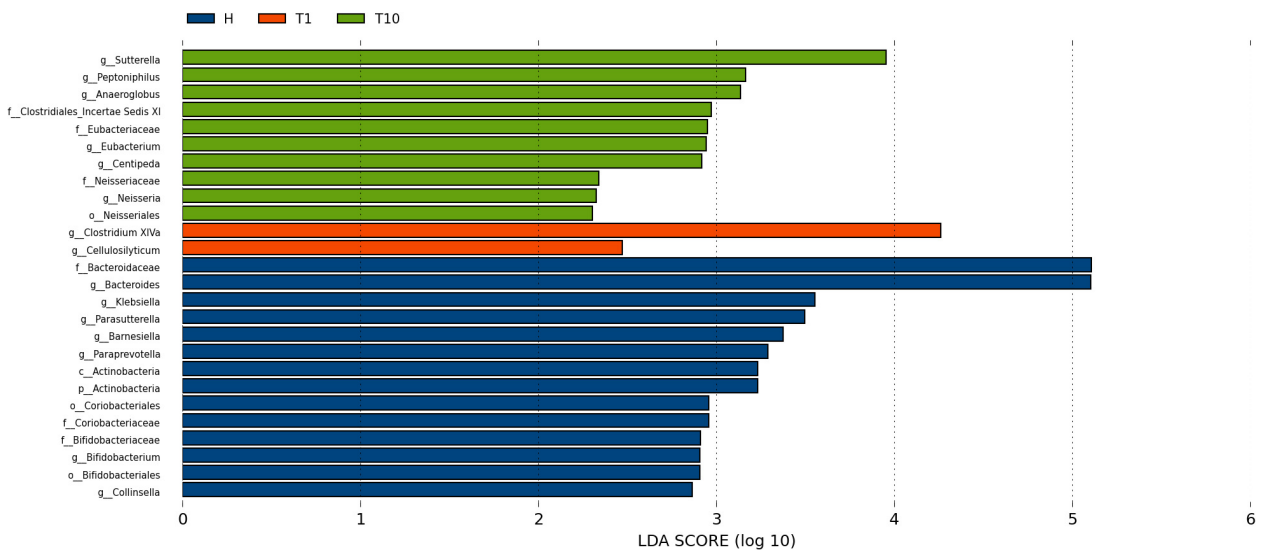


FIGURE 3 | Microbial biomarkers among healthy individuals (H) and NLCLC patients (before the first therapy T1 and before the tenth therapy T10). Significantly different abundant taxa as biomarkers using Kruskal–Wallis test ($P < 0.05$) with LDA score > 2.0. based on LEfSe analysis.

(2017), and Zhuang et al. (2019) found that patients with lung cancer in fecal microbiome showed elevated levels of Enterococcus, which could lead to increased DNA mismatch rate that indirectly promote rectal cancer. However, there were no significant differences in intestinal microbiota between NSCLC patients and healthy individuals and between pre-therapy and

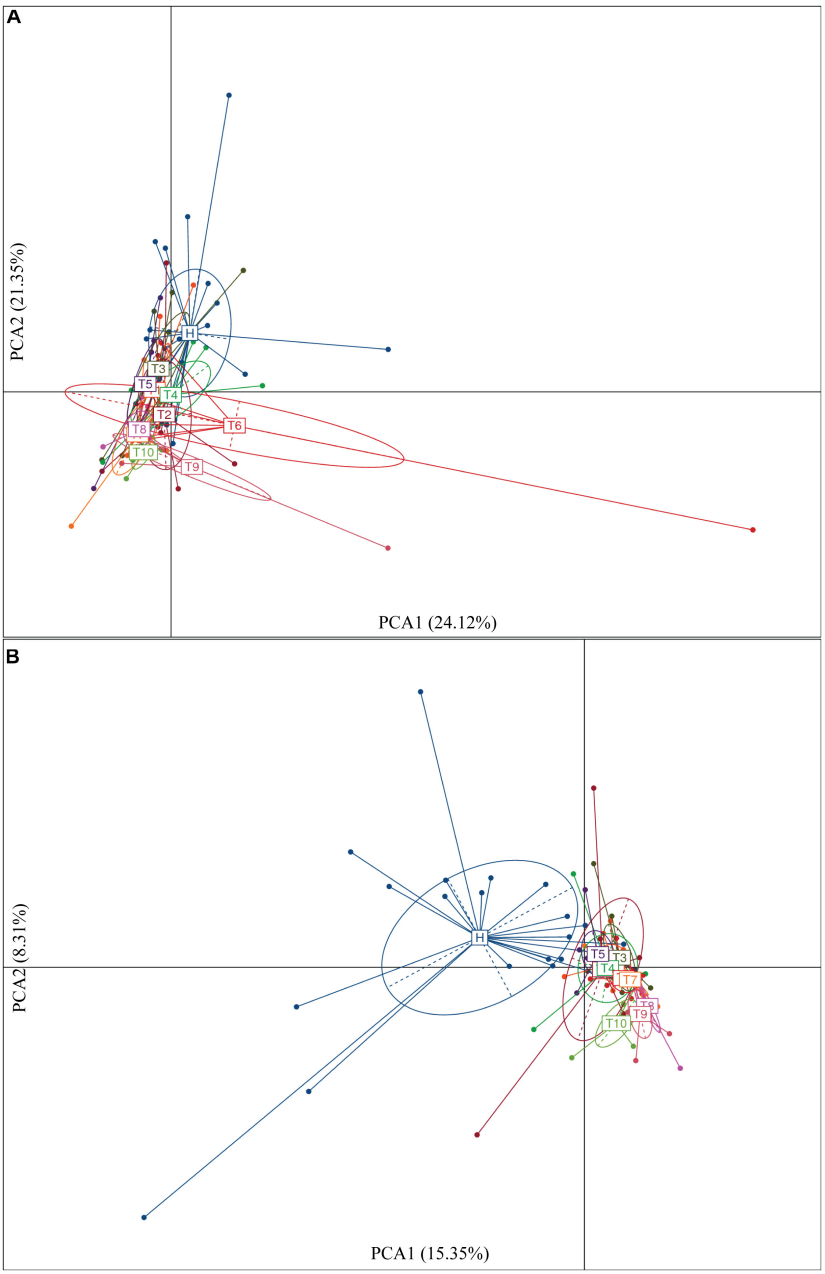


FIGURE 4 | Significantly different genera **(A)** and OTUs **(B)** based on the PCA plot between healthy individuals and NSCLC patients in response to osimertinib therapy.

post-therapy at the phylum, family, and genus level in our pilot study ($P > 0.05$, **Supplementary Tables S1–S3**). It suggested that osimertinib therapy did not greatly change the relative abundance of intestinal microbiota in NSCLC patients based on the taxonomical level, and that intestinal microbiota of these NSCLC patients at the baseline did not differ more than that of healthy individuals.

Secondly, we explored the differences of alpha and beta diversity of intestinal microbiota from NSCLC patients in response to osimertinib therapy. In previous study, Zhuang et al. (2019) found that there was no significant reduction in alpha diversity of intestinal microbiota in lung cancer patients compared to healthy individuals. In line with it, our results indicated that there were almost no significant differences in richness, phylogenetic diversity, and Shannon diversity between NSCLC patients and healthy individuals ($P > 0.05$, **Supplementary Table S4**). Moreover, no significant difference was observed in the richness and Shannon diversity of intestinal microbiota in NSCLC patients between pre-therapy (T1) and post-therapy (T2, T3, T4, T5, T6, T7, T8, T9, T10) ($P > 0.05$,

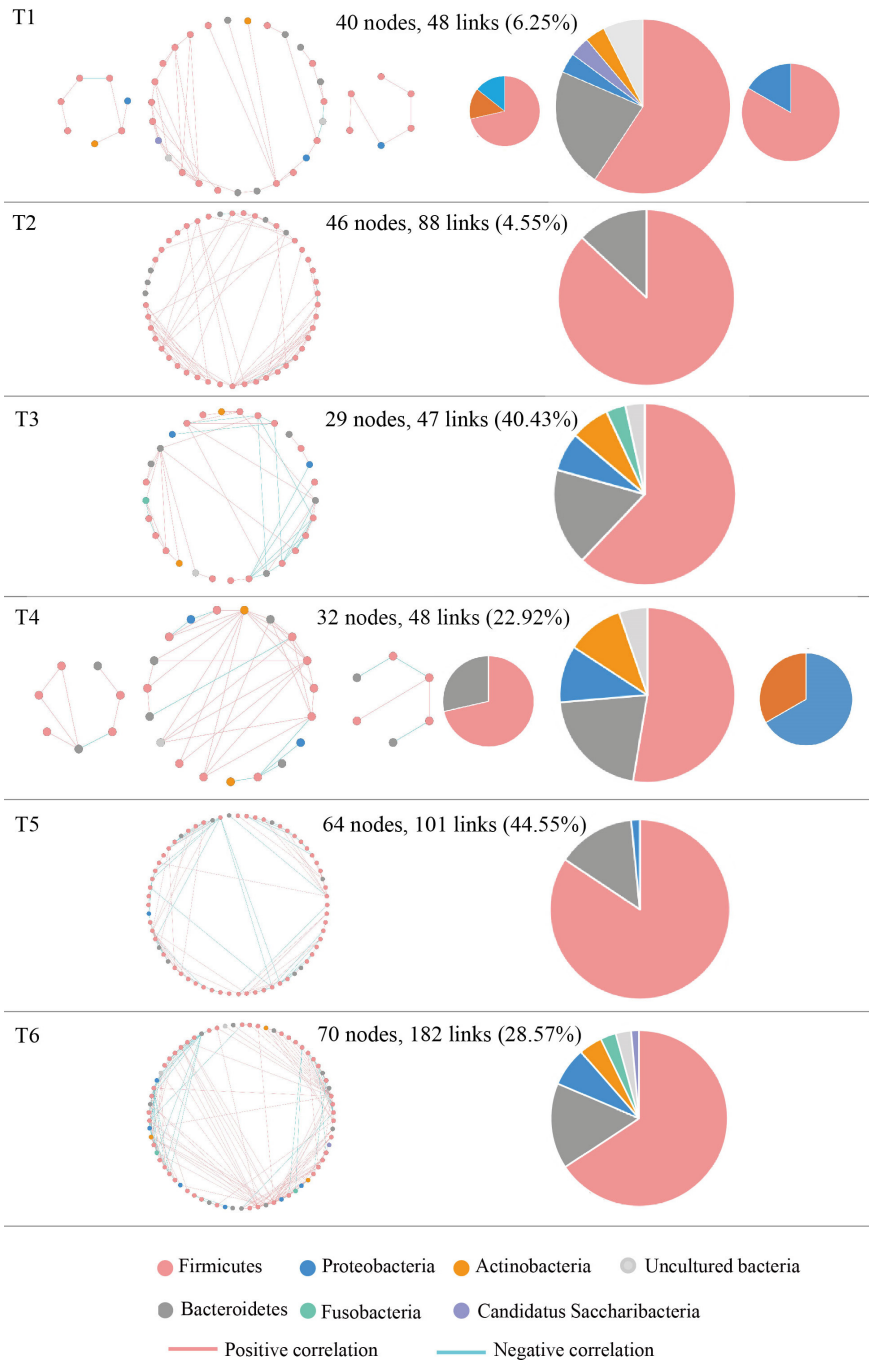


FIGURE 5 | Highly connected modules with intestinal microbial networks of NSCLC patients in response to osimertinib therapy. Node colors represent different phyla; pie charts represent the composition of the modules. A blue link indicates a negative relationship between two phyla, whereas a pink indicates a positive relationship. The number in bracket means the ratio of negative links accounting for the total links.

Supplementary Table S4), indicating that osimertinib therapy did not play great roles in alpha diversity of intestinal microbiota. However, we found that there was significantly different in beta diversity between pre-therapy and post-therapy in NSCLC patients based on the dissimilar test ($P < 0.05$, **Supplementary Table S5**), suggesting that osimertinib therapy has made the intestinal microbial community composition changed from the whole (Zhuang et al., 2019). At the same time, there was a little separation among NSCLC samples with different treatment cycles, and between healthy individuals and NSCLC patients (**Figure 4**). Previous studies reported that adaptive immunity in response to cancer therapy could shape the colonic microbiome

(Scholz et al., 2014). We speculated that the differences probably due to the different immune status and dietary behavior among them in response to osimertinib therapy.

Thirdly, we found the variations of the intestinal microbial markers in NSCLC patients before the first treatment (T1) and after the ninth treatment (T10). The microbial biomarkers in healthy individuals were the *Bacteroides*, *Klebsiella*, and *Parasutterella*. *Bacteroides*, commonly found in the human intestine, has a symbiotic host-bacterial relationship with humans. They assist in digesting food and producing valuable nutrients and energy to meet the body needs. Some strains of *Klebsiella* are considered as a part of the normal flora of the human gastrointestinal tract. The genus of *Parasutterella* has been defined as a core component of human and mouse gut microbiota, and has been correlated with various health outcomes (Ju et al., 2019). These indicated that our healthy samples were eligible. The genus *Clostridium XIVa* was considered as the biomarker in T1, which produce butyrate and other short chain fatty acids, has been correlated with susceptibility to enteric pathogens (Lopetuso et al., 2013). It indicated that intestinal state in NSCLC patients was relatively healthy before the osimertinib therapy. The *Sutterella*, *Peptoniphilus*, and *Anaeroglobus* dominated in T10. *Sutterella* spp. has been associated with autism, gastrointestinal dysfunction and metabolic syndrome (Williams et al., 2012; Lavelle et al., 2015; Lim et al., 2017). *Peptoniphilus* are important causes of bloodstream infection (Brown et al., 2014). *Anaeroglobus* as an opportunistic pathogen was reported in clinical infection that presented as pneumonia with empyema (Wang et al., 2015). Changes of intestinal microbial markers between pre-therapy and post-therapy showed that osimertinib therapy had certain effects on the biomarker microbes, suggesting that there were probably underlying intestinal problems.

Finally, MENs of intestinal microbiota in NSCLC patients were also changed in response to osimertinib therapy. Microbes in the intestine are not independent individuals; however, they always make intricate and inter-connected ecological communities. The links between nodes (taxa) could explain the co-exclusion or co-occurrence correlations, mainly caused by the species performing exclusive and complementary functions (Zhou et al., 2011). The study results showed that the links in the module were distinctly increased from T1 to T6, suggesting that intestinal microbial interspecies interactions within the constructed ecological networks were changed, and the more complicated and compact of module was made in response to osimertinib therapy (**Figure 5**). Positive interactions usually signify that nodes cooperate with one another, while negative interactions indicate competition between the taxa (Deng et al., 2012). Violle et al. (2010) established the protist communities in laboratory microcosms to demonstrate that external disturbance accelerate microbial species-species competition. In our pilot study, the negative links increased distinctly from T1 to T6, suggesting that osimertinib probably played key roles in the competition relationships based on the species-species interactions of intestinal microbiota (**Figure 5**).

Although we followed the longitudinal sampling of these NSCLC patients for about 1 year, there are a number of limitations in the present study. Since only 8 patients were enrolled in our study, data of more participants are needed. Furthermore, we only collected stool samples on the basis of administration. We will carry out further study to collect stool samples on the basis of dose and duration of administration of the drug. In the future, the detail therapeutic modalities and clinical settings in targeting the "gut-lung axis" need to be paid more attention for solving NSCLC that seems promise. In addition, since the main research object of intestinal microbial diversity analysis is intestinal bacteria, the whole process of the experiment was carried out using bacterial universal primers for the amplification of bacterial marker genes; the virus and mycoplasma present in a small part were not analyzed. It is necessary to design a completion plan for this limitation in the future study.

CONCLUSION

In conclusion, our pilot study found that osimertinib therapy changed intestinal microbial community composition from the whole, and made the intestinal microbial markers changed, as well as the varied microbial ecological networks for NSCLC patients. However, few roles were found in microbial composition changes at different taxonomical level and alpha diversity in response to osimertinib therapy. It indicated that osimertinib did not make radical change in intestinal microbiota of NSCLC patients. The partly changes of intestinal microbiota seem to be closely correlated with the few intestinal side effects and higher efficacy in these NSCLC patients with T790M mutation in response to osimertinib therapy.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Affiliated Hospital of Qingdao University Institutional Review Board. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

JC was involved in the conception and design of the study. JC, DL, CZ, MX, RZ, and HZ were involved in the collection and

assembly of data. JC, YZ, and YX were involved in the data analysis and interpretation. All authors interpreted the data and wrote the manuscript and approved the final manuscript.

FUNDING

This work was supported by the funding from research funding by the Qingdao University of Science and Technology (1203043003670).

ACKNOWLEDGMENTS

We thank the NSCLC patients and healthy volunteers for providing the fecal samples that were used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.583525/full#supplementary-material

REFERENCES

Alexander, J. L., Wilson, I. D., Teare, J., Marchesi, J. R., Nicholson, J. K., and Kinross, J. M. (2017). Gut microbiota modulation of chemotherapy efficacy and toxicity. *Nat. Rev. Gastroenterol. Hepatol.* 14, 356–365. doi: 10.1038/nrgastro.2017.20

Amarnani, R., and Rapose, A. (2017). Colon cancer and *Enterococcus bacteremia* co-affection: a dangerous alliance. *J. Infect. Public Health* 10, 681–684. doi: 10.1016/j.jiph.2016.09.009

Arrieta, M. C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., et al. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* 7:307ra152. doi: 10.1126/scitranslmed.aab2271

Brennan, C. A., and Garrett, W. S. (2019). *Fusobacterium nucleatum* - symbiont, opportunist and oncobacterium. *Nat. Rev. Microbiol.* 17, 156–166. doi: 10.1038/s41579-018-0129-6

Brown, K., Church, D., Lynch, T., and Gregson, D. (2014). Bloodstream infections due to *Peptoniphilus* spp.: report of 15 cases. *Clin. Microbiol. Infect.* 20, O857–O860.

Budden, K. F., Gellatly, S. L., Wood, D. L., Cooper, M. A., Morrison, M., Hugenholtz, P., et al. (2017). Emerging pathogenic links between microbiota and the gut-lung axis. *Nat. Rev. Microbiol.* 15, 55–63. doi: 10.1038/nrmicro.2016.142

Cong, J., Zhu, H., Liu, D., Li, T., Zhang, C., Zhu, J., et al. (2018). A Pilot study: changes of gut microbiota in post-surgery colorectal cancer patients. *Front. Microbiol.* 9:2777. doi: 10.3389/fmicb.2018.02777

Cong, J., Zhu, J., Zhang, C., Li, T., Liu, K., Liu, D., et al. (2019). Chemotherapy alters the phylogenetic molecular ecological networks of intestinal microbial communities. *Front. Microbiol.* 10:1008. doi: 10.3389/fmicb.2018.1008

Cross, D. A. E., Ashton, S. E., Serban, G., Cath, E., Nebhan, C. A., Spitzler, P. J., et al. (2014). AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. *Cancer Discov.* 4, 1046–1061. doi: 10.1158/2159-8290.cd-14-0337

Deng, Y., Jiang, Y. H., Yang, Y., He, Z., Luo, F., and Zhou, J. (2012). Molecular ecological network analyses. *BMC Bioinform.* 13:113. doi: 10.1186/1471-2105-13-113

Dickson, R. P., Martinez, F. J., and Huffnagle, G. B. (2014). The role of the microbiome in exacerbations of chronic lung diseases. *Lancet* 384, 691–702. doi: 10.1016/s0140-6736(14)61136-3

Doestzada, M., Vila, A. V., Zhernakova, A., Koonen, D. P. Y., Weersma, R. K., Touw, D. J., et al. (2018). Pharmacomicrobiomics:anovelroutetowardspersonalizedmedicine?. *Protein Cell* 9, 432–445. doi: 10.1007/s13238-018-0547-2

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Garzoni, C., Brugger, S. D., Qi, W., Wasmer, S., Cusini, A., Dumont, P., et al. (2013). Microbial communities in the respiratory tract of patients with interstitial lung disease. *Thorax* 68, 1150–1156. doi: 10.1136/thoraxjnl-2012-202917

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. (2007). Classes of complex networks defined by role-to-role connectivity profiles. *Nat. Phys.* 3, 63–69. doi: 10.1038/nphys489

Haiser, H. J., Gootenberg, D. B., Chatman, K., Sirasani, G., Balskus, E. P., and Turnbaugh, P. J. (2013). Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science* 341, 295–298. doi: 10.1126/science.1235872

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Ju, T., Kong, J. Y., Stothard, P., and Willing, B. P. (2019). Defining the role of parasutterella, a previously uncharacterized member of the core gut microbiota. *ISME J.* 13, 1520–1534. doi: 10.1038/s41396-019-0364-5

Kong, Y. (2011). Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 98, 152–153. doi: 10.1016/j.ygeno.2011.05.009

Lavelle, A., Lennon, G., O'sullivan, O., Docherty, N., Balfe, A., Maguire, A., et al. (2015). Spatial variation of the colonic microbiota in patients with ulcerative colitis and control volunteers. *Gut* 64, 1553–1561. doi: 10.1136/gutjnl-2014-307873

Lim, M. Y., You, H. J., Yoon, H. S., Kwon, B., Lee, J. Y., Lee, S., et al. (2017). The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. *Gut* 66, 1031–1038. doi: 10.1136/gutjnl-2015-311326

Lopetuso, L. R., Scaldaferri, F., Petito, V., and Gasbarrini, A. (2013). Commensal Clostridia: leading players in the maintenance of gut homeostasis. *Gut Pathog.* 5:23. doi: 10.1186/1757-4749-5-23

Magoc, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Mok, T. S., Wu, Y. L., Ahn, M. J., Garassino, M. C., Kim, H. R., Ramalingam, S. S., et al. (2017). Osimertinib or platinum-pemetrexed in EGFR T790M-positive lung cancer. *N. Engl. J. Med.* 376, 629–640.

Rajilic-Stojanovic, M., and De Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol. Rev.* 38, 996–1047. doi: 10.1111/1574-6976.12075

Rajpoot, M., Sharma, A. K., Sharma, A., and Gupta, G. K. (2018). Understanding the microbiome: emerging biomarkers for exploiting the microbiota for personalized medicine against cancer. *Semin. Cancer Biol.* 52, 1–8. doi: 10.1016/j.semcancer.2018.02.003

Ramalingam, S. S., Vansteenkiste, J., Planchard, D., Cho, B. C., Gray, J. E., Ohe, Y., et al. (2020). Overall survival with osimertinib in untreated, EGFR-mutated advanced NSCLC. *N. Engl. J. Med.* 382, 41–50. doi: 10.1056/nejmoa1913662

Rosell, R., and Karachaliou, N. (2016). Large-scale screening for somatic mutations in lung cancer. *Lancet* 387, 1354–1356. doi: 10.1016/s0140-6736(15)01125-3

Santarpia, M., Karachaliou, N., and Rosell, R. (2017a). Beyond platinum treatment for NSCLC: what does the future hold? *Expert. Rev. Anticancer Ther.* 17, 293–295. doi: 10.1080/14737140.2017.1288103

Santarpia, M., Liguori, A., Karachaliou, N., Gonzalezcao, M., Daffinà, M. G., D'aveni, A., et al. (2017b). Osimertinib in the treatment of non-small-cell lung cancer: design, development and place in therapy. *Lung Cancer Targ. Ther.* 8, 109–125. doi: 10.2147/lctt.s119644

Scholz, F., Badgley, B. D., Sadowsky, M. J., and Kaplan, D. H. (2014). Immune mediated shaping of microflora community composition depends on barrier site. *PLoS One* 9:e084019. doi: 10.1371/journal.pone.0084019

Schwartz, L. H., Litière, S., De Vries, E., Ford, R., Gwyther, S., Mandrekar, S., et al. (2016). RECIST 1.1-Update and clarification: from the RECIST committee. *Eur. J. Cancer* 62, 132–137. doi: 10.1016/j.ejca.2016.03.081

Soria, J. C., Ohe, Y., Vansteenkiste, J., Reungwetwattana, T., Chewaskulyong, B., Lee, K. H., et al. (2018). Osimertinib in untreated EGFR-mutated advanced non-small-cell lung cancer. *N. Engl. J. Med.* 378, 113–125.

Strickertsson, J. A., Desler, C., Martin-Bertelsen, T., Machado, A. M., Wadstrøm, T., Winther, O., et al. (2013). Enterococcus faecalis infection causes inflammation, intracellular oxphos-independent ROS production, and DNA damage in human gastric cancer cells. *PLoS One* 8:e63147. doi: 10.1371/journal.pone.0063147

Tiago, S., Vipul, Y., Vanessa, Z., Anders, B., Bertil, A., and Basit, A. W. (2014). On the colonic bacterial metabolism of azo-bonded prodrugsof 5-aminosalicylic acid. *J. Pharm. Sci.* 103, 3171–3175. doi: 10.1002/jps.24103

Torre, L. A., Siegel, R. L., Ward, E. M., and Jemal, A. (2016). Global cancer incidence and mortality rates and trends–an update. *Cancer Epidemiol. Biomark. Prev.* 25, 16–27. doi: 10.1158/1055-9965.epi-15-0578

Vael, C., Nelen, V., Verhulst, S. L., Goossens, H., and Desager, K. N. (2008). Early intestinal *Bacteroides fragilis* colonisation and development of asthma. *BMC Pulm. Med.* 8:19. doi: 10.1186/1471-2105-13-19

Violle, C., Pu, Z., and Jiang, L. (2010). Experimental demonstration of the importance of competition under disturbance. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12925–12929. doi: 10.1073/pnas.1000699107

Wallace, B. D., Wang, H., Lane, K. T., Scott, J. E., Orans, J., Koo, J. S., et al. (2010). Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science* 330, 831–835. doi: 10.1126/science.1191175

Wang, C. H., Kan, L. P., Sun, J. R., Yu, C. M., Yin, T., Huang, T. W., et al. (2015). Empyema caused by Anaeroglobus geminates, a case report with literature review. *Infection* 43, 117–120. doi: 10.1007/s15010-014-0679-0

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/aem.00062-07

Wexler, A. G., and Goodman, A. L. (2017). An insider's perspective: *Bacteroides* as a window into the microbiome. *Nat. Microbiol.* 2:17026.

Williams, B. L., Mady, H., Tanmay, P., and Ian, W. L. (2012). Application of novel PCR-based methods for detection, quantitation, and phylogenetic characterization of *Sutterella* species in intestinal biopsy samples from children with autism and gastrointestinal disturbances. *mBio* 3, 54–64.

Wu, H., Esteve, E., Tremaroli, V., Khan, M. T., Caesar, R., Mannerås-Holm, L., et al. (2017). Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med.* 23:850. doi: 10.1038/nm.4345

Wu, X., Zhang, H., Chen, J., Shang, S., Wei, Q., Yan, J., et al. (2016). Comparison of the fecal microbiota of dholes high-throughput Illumina sequencing of the V3-V4 region of the 16S rRNA gene. *Appl. Microbiol. Biotechnol.* 100, 3577–3586. doi: 10.1007/s00253-015-7257-y

Zhou, J., Deng, Y., Luo, F., He, Z., and Yang, Y. (2011). Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO_2. *mBio* 2:e00122-11.

Zhuang, H., Cheng, L., Wang, Y., Zhang, Y. K., Zhao, M. F., Liang, G. D., et al. (2019). Dysbiosis of the gut microbiome in lung cancer. *Front. Cell Infect. Microbiol.* 9:112. doi: 10.3389/fmicb.2018.00112

Zimmermann, M., and Zimmermann-Kogadeeva, M. (2019). Separating host and microbiome contributions to drug pharmacokinetics and toxicity. *Science* 363:eaat9931. doi: 10.1126/science.aat9931

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 12 November 2020
doi: 10.3389/fmicb.2020.570825

Data-Driven Modeling for Species-Level Taxonomic Assignment From 16S rRNA: Application to Human Microbiomes

*Ho-Jin Gwak[1] and Mina Rho[1,2]**

[1] Department of Computer Science and Engineering, Hanyang University, Seoul, South Korea, [2] Department of Biomedical Informatics, Hanyang University, Seoul, South Korea

With the emergence of next-generation sequencing (NGS) technology, there have been a large number of metagenomic studies that estimated the bacterial composition via 16S ribosomal RNA (16S rRNA) amplicon sequencing. In particular, subsets of the hypervariable regions in 16S rRNA, such as V1–V2 and V3–V4, are targeted using high-throughput sequencing. The sequences from different taxa are assigned to a specific taxon based on the sequence homology. Since such sequences are highly homologous or identical between species in the same genus, it is challenging to determine the exact species using 16S rRNA sequences only. Therefore, in this study, *homologous species groups* were defined to obtain maximum resolution related with species using 16S rRNA. For the taxonomic assignment using 16S rRNA, three major 16S rRNA databases are independently used since the lineage of certain bacteria is not consistent among these databases. On the basis of the NCBI taxonomy classification, we re-annotated inconsistent lineage information in three major 16S rRNA databases. For each species, we constructed a consensus sequence model for each hypervariable region and determined *homologous species groups* that consist of indistinguishable species in terms of sequence homology. Using a *k*-nearest neighbor method and the species consensus sequence models, the species-level taxonomy was determined. If the species determined is a member of *homologous species groups*, the species group is assigned instead of a specific species. Notably, the results of the evaluation on our method using simulated and mock datasets showed a high correlation with the real bacterial composition. Furthermore, in the analysis of real microbiome samples, such as salivary and gut microbiome samples, our method successfully performed species-level profiling and identified differences in the bacterial composition between different phenotypic groups.

Keywords: 16S rRNA, microbial community, differential composition, operational taxonomic units, taxonomy assignment

OPEN ACCESS

Edited by:
Steve Lindemann,
Purdue University, United States

Reviewed by:
Tsute Chen,
The Forsyth Institute, United States
Martin W. Hahn,
University of Innsbruck, Austria

***Correspondence:**
Mina Rho
minarho@hanyang.ac.kr

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *09 June 2020*
Accepted: *22 October 2020*
Published: *12 November 2020*

Citation:
Gwak H-J and Rho M (2020) Data-Driven Modeling for Species-Level Taxonomic Assignment From 16S rRNA: Application to Human Microbiomes. Front. Microbiol. 11:570825. doi: 10.3389/fmicb.2020.570825

INTRODUCTION

Metagenomics has been widely used to analyze microbial communities without cultivating strains (Breitbart et al., 2003; Schloss and Handelsman, 2003; Handelsman, 2004; Petrosino et al., 2009; Qin et al., 2010; Peng et al., 2019; Yang L. et al., 2019; Brumfield et al., 2020; Chung et al., 2020; Khachatryan et al., 2020). Moreover, the 16S ribosomal RNA (16S rRNA) gene has been regarded as an informative resource for the identification of the species and the estimation of bacterial composition as it has both well-conserved and hypervariable regions among different species. Thus, the conserved regions can be used as primers to target specific hypervariable regions using targeted amplicon sequencing (Petrosino et al., 2009), whereas the hypervariable regions can be used to identify bacterial taxonomy using the sequence similarities between different species. Although the 16S rRNA gene is a useful material to identify bacteria, it is challenging to completely discriminate species since 16S rRNA genes are identical or highly homologous between some different species. Genome comparisons by DNA–DNA hybridization or genome sequence comparison (ANI analyses) were needed to assign an exact species (Cho and Tiedje, 2001; Ciufo et al., 2018).

Using 454 pyrosequencing (Petrosino et al., 2009; Cummings et al., 2013) and Illumina MiSeq technology (Wen et al., 2017; Ravi et al., 2018; Sessou et al., 2019), 16S rRNA analysis pipelines were built to estimate the bacterial composition of different species (Turnbaugh et al., 2007; Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012). While attempts are being made to analyze the entire 16S rRNA sequence via long-read sequencing using PacBio (Quail et al., 2012) or Oxford Nanopore (Winand et al., 2019) technology, the high error rates and costs limit their practical utility. When estimating bacterial composition using targeted amplicon sequencing, the results might differ depending on the choice of hypervariable regions, such as V1–V2 or V3–V4. Therefore, selecting appropriate hypervariable regions for analysis is important. Several studies have been conducted to investigate the manner in which the analysis of different variable regions affects the estimation of bacterial composition (Sun et al., 2013; Johnson et al., 2019).

The 16S rRNA analysis pipeline involves preprocessing, clustering [operational taxonomic units (OTU) picking], assigning taxonomy, and estimating the bacterial composition. Although most of the sequencing errors are filtered out at the preprocessing step, there are still some sequencing errors that remain. To overcome these errors and strain variations, processed reads are clustered into OTUs using a 97 or 99% sequence similarity threshold. Since sequences belonging to the same OTU are considered to be derived from the same clade, OTU clustering directly affects the estimation of bacterial composition. Therefore, several clustering algorithms have been developed to overcome strain variation and sequencing errors. For example, the UPARSE algorithm (Edgar, 2013) clusters sequences on the basis of sequence similarity, whereas the Minimum Entropy Decomposition (MED) (Eren et al., 2015) and DADA2 (Callahan et al., 2016) algorithms cluster sequences via the association of position-specific variations. For taxonomy assignment, classifiers such as MEGAN (Huson et al., 2007), RDP naïve Bayesian classifier (Wang et al., 2007), Kraken (Wood and Salzberg, 2014), and SPINGO (Allard et al., 2015) were developed. Thus, not only the classifier but also the 16S rRNA database is important for accurate taxonomical classification. There are currently three major 16S rRNA databases that are widely used, namely GreenGenes (DeSantis et al., 2006), SILVA (Quast et al., 2013), and RDP (Cole et al., 2014). However, although new bacterial taxa continue to be reported, these three databases have not been updated for over 2 years. Furthermore, the lineage of some bacteria is not consistent among these three databases (Balvociute and Huson, 2017; Edgar, 2018a).

In this study, we re-annotated the inconsistent or mislabeled taxa in the three 16S rRNA databases on the basis of the NCBI taxonomy classification. The 16S rRNA sequences were combined from the re-annotated GreenGenes, SILVA, and NCBI databases to include species that exist exclusively in each database or were recently annotated. In the evaluation of taxonomy classification, the classifier trained with all three databases showed the best accuracy in terms of precision and recall rates. Moreover, *taxonomic separability* was measured for the V1–V2 and V3–V4 hypervariable regions at the genus and species level. For each species, we constructed consensus sequences for each hypervariable region and determined indistinguishable species. By comparing the consensus sequences of each species, *homologous species groups* in which the species share high similarity were constructed for each hypervariable region, which was then used for the species-level taxonomy assignment. The evaluation performed using simulated datasets and mock datasets showed a high correlation with the real bacterial composition. Moreover, when analyzing real microbiomes, such as the salivary and gut microbiome, our method successfully performed species-level profiling to identify differences in bacterial composition between different phenotypic groups.

MATERIALS AND METHODS

Re-annotating the 16S rRNA Sequence Databases

To investigate the taxonomy consistency, GreenGenes v13.5, SILVA v132, and RDP v11.5 databases were used. As quality control, sequences whose length range in three times the standard deviation from the mean without any ambiguous nucleotide (e.g., N) were used. Out of 1,242,330, 1,861,373, and 3,196,041 sequences obtained from GreenGenes, SILVA, and RDP databases, respectively, 1,191,315, 1,779,305 and 1,559,121 sequences were retained for the re-annotation after quality control process (**Supplementary Table 1**).

To apply the latest version of NCBI taxonomy, NCBITaxa class in the ete3 python package (Federhen, 2012; Huerta-Cepas et al., 2016) was used with NCBI taxdump downloaded on January 3, 2020. The taxonomy tree with seven taxonomic ranks (superkingdom, phylum, class, order, family, genus, and species) was used in this re-annotation.

Each 16S rRNA sequence was re-annotated using the taxon at the lowest taxonomy rank in the database (**Figure 1** and

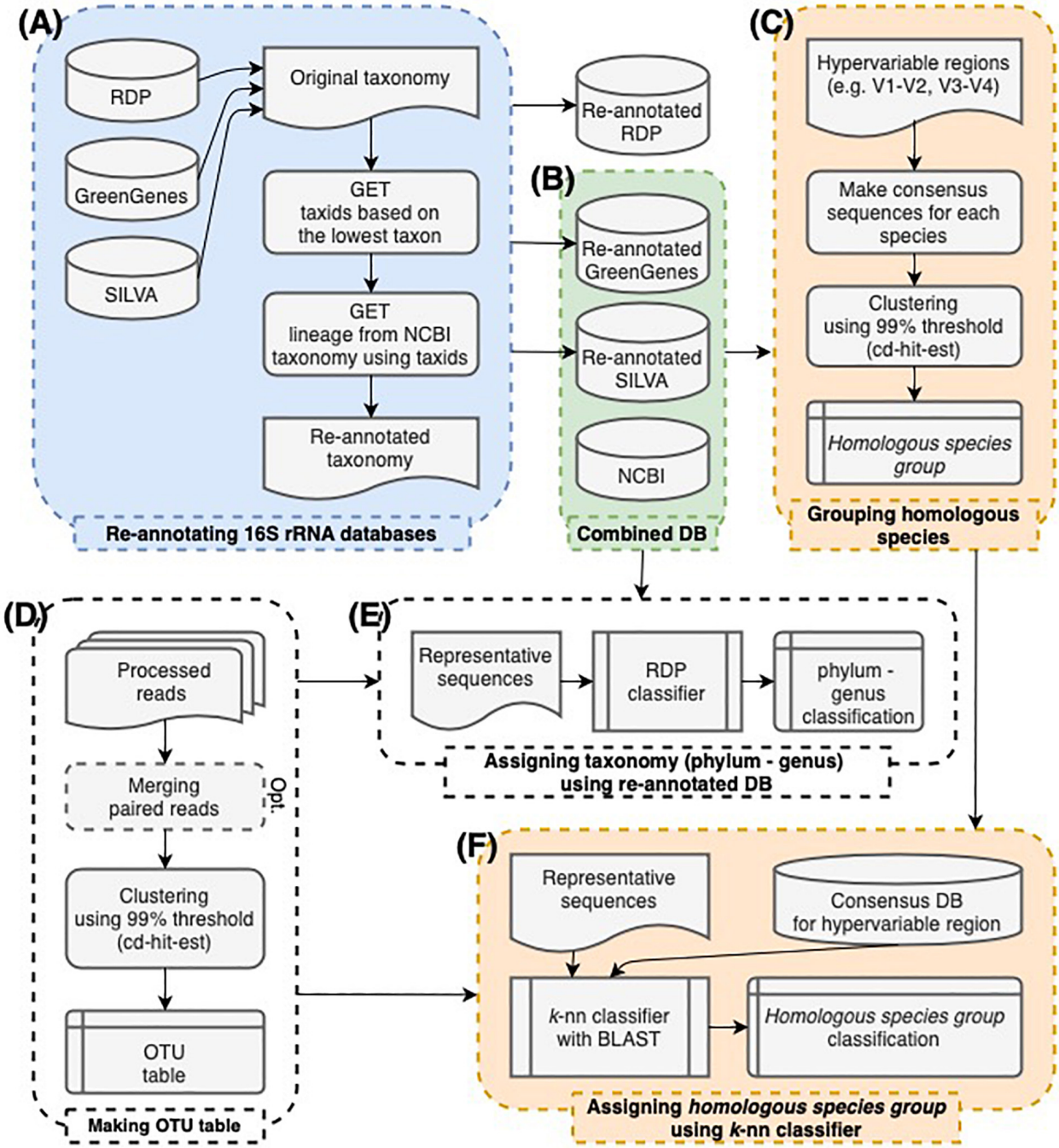


FIGURE 1 | The workflow of this study. **(A)** The taxonomy of 16S rRNA sequences in GreenGenes, SILVA, and RDP databases was re-annotated according to the NCBI taxonomy. **(B)** GreenGenes, SILVA, and NCBI databases were combined. RDP database was excluded since they have no species annotation. **(C)** Consensus sequences of each species were made by hypervariable region sequences extracted from the combined databases. By clustering consensus sequences within 99% sequence similarity, *homologous species groups* were established. **(D)** OTU tables were made in a conventional manner. **(E)** The taxonomy assignment from phylum to genus-level was processed by the RDP naïve Bayesian classifier re-trained using the combined database. **(F)** The species-level classification was processed by searching sequences against the consensus database. Sequences were labeled as the representative species of the *homologous species group* that includes the best hit of the sequence.

Supplementary Figure 1). To identify the lowest rank, the provided taxa were searched from species to superkingdom. For each rank, the taxid was returned if it was found using the get_name_translator() function in NCBITaxa class. Otherwise, that rank was skipped. When the species name was specified with the strain name at the species rank, only the species name

was used. Since *Escherichia* and *Shigella* species have essentially identical 16S rRNA sequences, sequences labeled as *Escherichia* or *Shigella* were collectively labeled as *Escherichia.Shigella*.

Gathering the 16S rRNA Sequences From the Genomes in the NCBI RefSeq Database

The genomes assembled at the complete-level or chromosome-level were downloaded from the ftp site of the NCBI RefSeq database[1]. The information for each genome is listed in the *assembly_summary_refseq.txt* file (downloaded on July 16, 2019). In the generic feature file (GFF), the regions where the feature is described as "rRNA" and the product as "16S ribosomal RNA" were identified as the 16S rRNA sequences and extracted from the genome. Thus, we obtained 78,270 16S rRNA sequences from the 16,337 genomes analyzed. As quality control, the same filtering step was performed for the sequences extracted and 77,803 16S rRNA sequences were retained to train the classifier and generate consensus models.

Simulating the Hypervariable Regions From the 16S rRNA Sequences

The 27F/308R and 337F/806R primer pairs are widely used to target the hypervariable regions V1–V2 and V3–V4, respectively, for Illumina MiSeq amplicon sequencing. The fragment sequences of the hypervariable regions were simulated by extracting the sequences between the forward and backward primers from the 16S rRNA sequences using cutadapt (Martin, 2011). Moreover, an error level of 20% (i.e., 2–3 nt mismatches) was allowed when matching the primer sequences. The mean and standard deviation of the extracted fragment length were also calculated. Fragments longer or shorter than twice the standard deviation from the mean value were ignored. Fragments containing "N" were also ignored (**Supplementary Table 2**).

Constructing *Homologous Species Groups* for Each Hypervariable Region

To determine which species are distinguishable by their 16S rRNA sequences, sequence similarities between species belonging to the same genus were calculated. A consensus sequence of the strains belonging to the same species was obtained using the "cons" function in EMBOSS v6.6.0 (Olson, 2002) with the default parameter settings. Pairwise sequence similarities were measured between the consensus sequences of each pair of species using the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) implemented in the "needle" function in EMBOSS v6.6.0 with the default parameter settings.

On the basis of the sequence similarity of the consensus sequences, *homologous species groups* that shared 99% or higher sequence similarity were constructed. The species in a *homologous species group* were considered indistinguishable by their 16S rRNA sequences. To name the *homologous species group*, the species in the group with the largest number of strains were selected and extended with a "+" sign.

[1] ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/

Simulating Amplicon Sequences From the Bacterial Genomes

Amplicon sequences for the V3–V4 hypervariable region were simulated from the bacterial reference genomes using Grinder (Angly et al., 2012). To target the V3–V4 hypervariable region, the 337F (CCTACGGGAGGCWGCAG) and 806R (GACTACHVGGGTMTCTAAT) primer sequences were used. For abundance models, the uniform, linear, and power-law with parameter 1 and 2 models were used. Amplicon sequences were simulated with a uniform 0.5% error model (-md uniform 0.5) and a length distribution of 421 ± 11 bp (-rd 421 uniform 11). Only the forward strands were used (-un 1), and the coverage fold was set to 1,000 (-cf 1000). Moreover, we considered copy number bias but not genome length bias (-cb 1 and -lb 0). All other parameters (i.e., those not mentioned above) were set as default.

Preprocessing of the Illumina Amplicon Sequencing Reads

The 16S rRNA genes were sequenced using the Illumina MiSeq sequencer, and paired-end reads were generated and merged on the basis of their overlapping region. Each read pair was assembled using FLASH (Magoc and Salzberg, 2011) with the default parameter settings except for a minimum overlap of 20 bp (-m 20) and maximum overlap of 300 bp (-M 300). Assembled contigs (including "N") were removed using an in-house script. Merged fragments longer than twice the standard deviation from the mean of the hypervariable region length (mean and standard deviation of the V3–V4 region were 421 and 11 nt, respectively) were also removed using Sickle. The mean and standard deviation of the V3–V4 region length were calculated from the sequences in the GreenGenes database.

Constructing the OTUs and Determining Their Taxonomy Assignment

The classification of the 16S rRNA sequence was performed according to the conventional classification approach (**Figure 1**). Preprocessed reads were clustered into OTUs using cd-hit-est (Fu et al., 2012). Cd-hit-est was used with the following parameter settings: no memory limitation (-M 0), word size 10 (-n 10), cluster into the most similar cluster (-g 1), and a 99% sequence similarity threshold (-c 0.99). The other parameters were set as default. Each representative sequence was classified using the RDP naïve Bayesian classifier trained with our combined database.

RESULTS AND DISCUSSION

Refinement of Inconsistent Taxonomy Annotation in 16S rRNA Databases

Using the 16S rRNA sequences from three major 16S rRNA databases, we investigated the consistency of the taxonomic lineage annotation. When we compared the taxonomic lineage annotations provided by the three 16S rRNA databases, we found that the same genus or species was often annotated with a different lineage. Out of the 1,122, 4,985, and 2,191 genera

included in the GreenGenes v.13.5, SILVA v.132, and RDP v.11.5 databases filtered, respectively, 183, 2,794, and 68 were exclusive to each database (**Supplementary Figure 2A**). Notably, out of the 853 genera included in all three databases, only 288 were annotated with the same lineage. Moreover, 112 genera were annotated with different lineages in all three databases. For example, the order of *Mycobacterium* was annotated as Actinomycetales in the GreenGenes and RDP databases but as Corynebacteriales in the SILVA database. The order of *Corynebacterium* was also annotated as Actinomycetales in the GreenGenes and RDP database but as Corynebacteriales in the NCBI taxonomy classification. Taxonomy reclassification also resulted in inconsistent taxonomic lineage annotation among the three databases. For example, *Propionibacterium* was originally identified as *Bacillus* but was later renamed as *Propionibacterium* (Douglas and Gunter, 1946). However, it was recently reclassified as *Cutibacterium* (Dreno et al., 2018).

We re-annotated the three major 16S rRNA databases based on NCBI taxonomy since inconsistencies between the databases could produce different bacterial composition profiles depending on the choice of database. Using the sequences filtered from GreenGenes, SILVA, and RDP databases, 667,528 (56%), 907,944 (51%), and 1,275,668 (82%) sequences were re-annotated in this study, respectively (**Supplementary Table 1**). As a result, we obtained 879 genera with the same lineage annotation among three databases (**Table 1**, **Supplementary Figure 2B**), compared to the 288 genera identified before the refinement step (**Supplementary Figure 2A**). Only four genera existed exclusively in the GreenGenes, 15 genera in the RDP, and 955 genera in the SILVA database (**Supplementary Figure 2B**).

After the re-annotation, the sequences from GreenGenes and SILVA databases were used in our classification method (**Figure 1B**). The RDP database was excluded since species-level annotation was not provided. In addition, the 16S rRNA sequences extracted from the complete genomes in the NCBI RefSeq database were included. In total, 823,937, 1,306,532, and 77,410 sequences with the genus-level annotation from GreenGenes, SILVA, and NCBI, respectively, were used in our classification method (**Supplementary Table 1**).

Genus-Level Profiling Using the Combined Database

For the genus-level taxonomy assignment, the RDP naïve Bayesian classifier was retrained with the sequences re-annotated in this study. Classifiers were tested using the V3–V4 region sequences extracted from the NCBI database. In the evaluation, the classifier trained with our combined database showed the best performance in terms of precision and recall rates from the phylum to genus level (**Figure 2** and **Table 2**). Notably, the classifier trained with one database (i.e., GreenGenes) had precision and recall rates of 89.33 and 81.85%, respectively, whereas the classifier trained with all three databases had precision and recall rates of 97.88 and 96.39%, respectively.

To evaluate the classification performance for the newly annotated bacteria, the gut microbiome of mice (Chung et al., 2020) were profiled using the classifier trained with our combined database (**Supplementary Figure 4**). In the previous report, the profiling of the bacterial composition of the samples using metagenomic reads revealed that Muribaculaceae and its genera were the most abundant taxon, whereas profiling via the 16S rRNA amplicon sequencing reads showed that Barnesiellaceae was the most abundant. This difference was explained by the different versions of the databases used (Chung et al., 2020). Sequences annotated with these two genera were not included in the GreenGenes database and RDP database, since *Muribaculum* and *Duncaniella* were first reported in the NCBI repository in July 2016 and March 2018, respectively. Notably, the classifier trained in our study correctly predicted the sequences as Muribaculaceae at the family rank, suggesting that the relative abundance of *Duncaniella* is similar to that obtained via metagenomic analysis. Although *Duncaniella* was well classified, *Muribaculum* was still reported with a low confidence score. This result suggests that there might be some genera belonging to the Muribaculaceae family that are still unknown.

TABLE 1 | The number of taxa for each taxonomic rank after re-annotation.

	Green genes	SILVA	RDP	NCBI
Superkingdom	1	1	1	1
Phylum	40	62	49	40
Class	72	85	73	72
Order	163	210	173	164
Family	355	495	380	358
Genus	1,030	3,239	2,154	1,206
Species	570	15,335	0	3,029
Number of sequences	1,191,315	1,779,305	1,559,121	77,869

Species-Level Profiling Using *Homologous Species Groups*

In conventional microbiota profiling, reads are clustered into OTUs based on sequence similarity. Most OTUs are created using a 97 or 99% sequence similarity threshold. These thresholds are based on the empirical observation of 94% or higher 16S rRNA sequence similarity within a genus and 97% or higher 16S rRNA sequence similarity within a species (Schloss and Handelsman, 2005). Note that, many studies have reported that species cannot be completely discriminated using such thresholds (Stackebrandt, 2006; Edgar, 2018b). We measured the *taxonomic separability* (i.e., how well different taxa are separately assigned to different OTUs) using the V1–V2 region, the V3–V4 region, and the entire 16S rRNA gene. OTUs were created using a 99% sequence similarity threshold to measure the proportion of OTUs that were assigned to multiple taxa (**Supplementary Figure 5**).

Most of the OTUs created consisted of sequences from one genus, whereas multiple species were assigned to the same OTU. Out of the 84,169, 127,223, and 179,039 OTUs created from the V1–V2 region, the V3–V4 region, and the entire 16S rRNA gene in the GreenGenes database, 3.58, 1.51, and 0.29% of the OTUs contained multiple species, respectively (**Supplementary Figure 5A**). In the SILVA database, out of the 118,404, 191,585, and 299,556 OTUs created, 20.26, 25.62, and 18.94% contained multiple species, respectively. Moreover, in the 16S rRNA gene

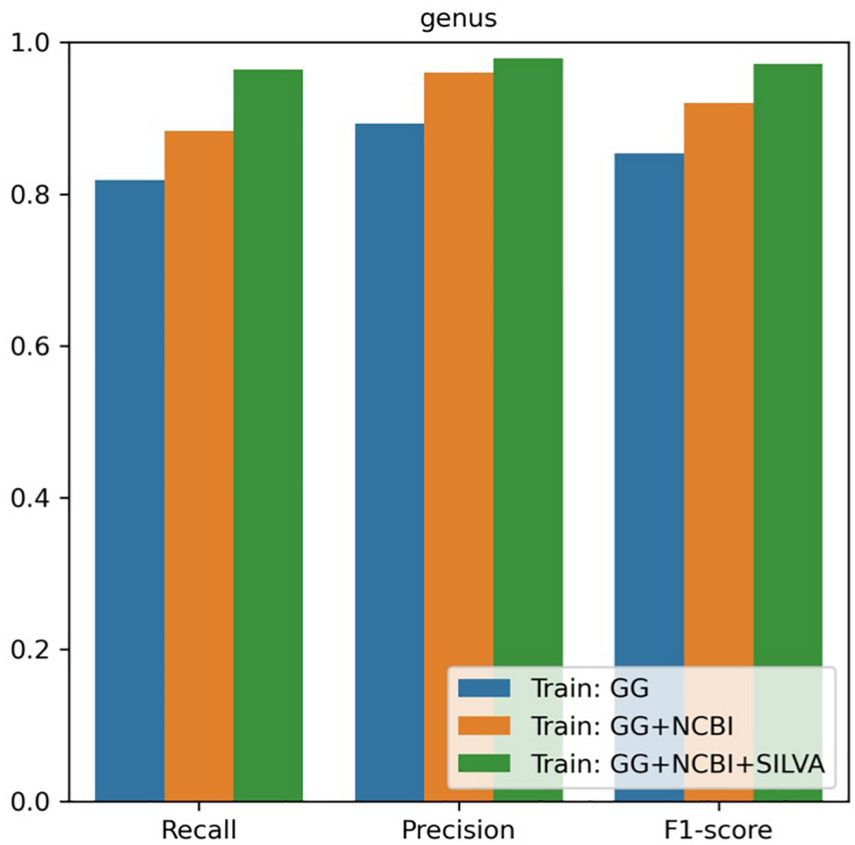


FIGURE 2 | Performance of the genus-level classification with the combined database.

sequences obtained from the NCBI database, 13.01, 19.54, and 13.44% of the 3,137, 2,746, and 3,987 OTUs created contained multiple species, respectively. While most of the sequences from different genera were assigned to different OTUs, almost half of the sequences from different species were assigned to the same OTU in the SILVA and the NCBI database (**Supplementary Figure 5B**). This result indicates that reads from such species are clustered together when OTUs are created using a 99% sequence similarity threshold.

To investigate *species separability* using the 16S rRNA sequences, a species network was constructed with the sequences of the V3–V4 region from our combined database (**Supplementary Figure 6**). In the network, each node is a consensus sequence of a species. If two species share 99% or higher sequence similarity, the nodes of those species were connected. Notably, many species from the same genus were clustered owing to the fact that their 16S rRNA sequences have 99% or higher similarity. Among the *Staphylococcus* species, seven groups were clustered, the largest of which consisted of 10 species (**Supplementary Figure 6A**). Moreover, 15 groups were clustered from the *Streptococcus* species, the largest of which consisted of eight species (**Supplementary Figure 6B**). The *homologous species groups* were constructed from the network analysis, which corresponded to the connected components in the graph.

TABLE 2 | Accuracy of the taxonomy classification when using the combined database.

	Recall	Precision	F1-score
Superkingdom	1	1	1
Phylum	0.9997	0.9995	0.99962
Class	0.9896	0.9989	0.99422
Order	0.9923	0.9981	0.99517
Family	0.9659	0.9954	0.98045
Genus	0.9832	0.9666	0.97482
Species	0.7696	0.7994	0.78423

In the *homologous species groups*, the consensus sequences of the included species had 99% or higher sequence similarity. **Figure 3** shows the *homologous species groups* in the arc of the same color, which resulted from the network analysis of two genera, *Staphylococcus* and *Streptococcus* (**Supplementary Figure 6**). Notably, some strain-level heterogeneity (i.e., 99% or higher sequence similarity between strains in different *homologous species groups*) was also observed (**Figure 3**). For example, some sequences belonging to *Staphylococcus aureus* and *Staphylococcus epidermidis* (labeled in blue and red) were connected. Such strain-level heterogeneity could be caused by either distinct strains in a specific species or incorrect annotation.

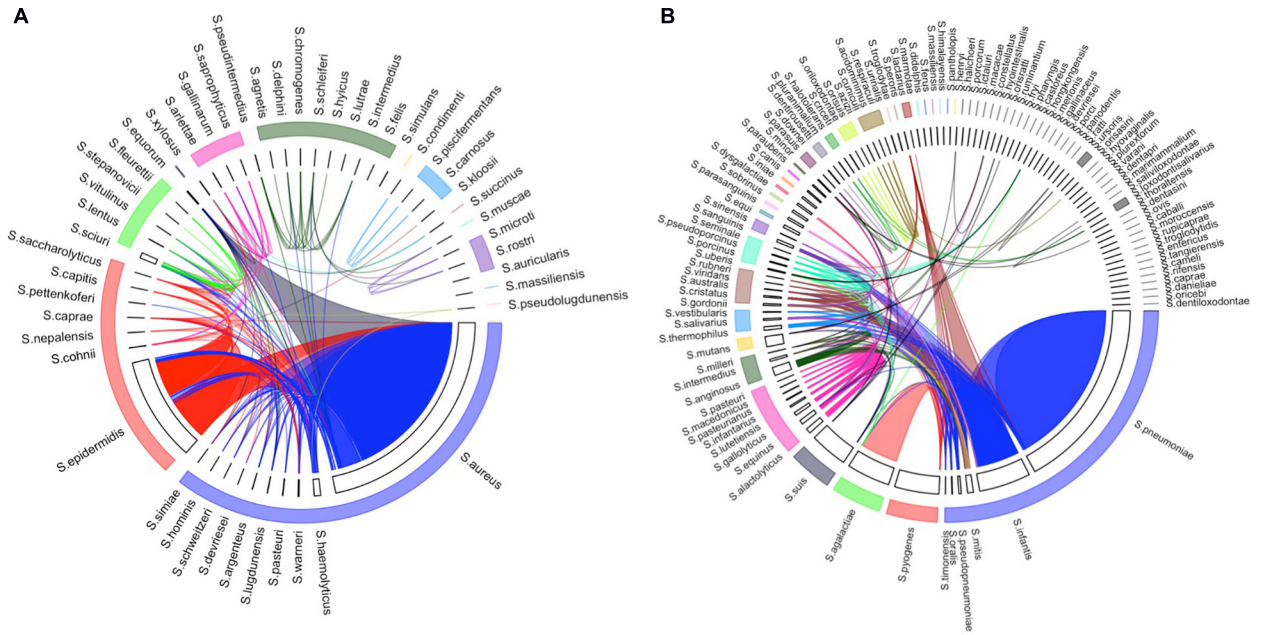


FIGURE 3 | *Homologous species groups* of the species in **(A)** *Staphylococcus* and **(B)** *Streptococcus*. The width of each track is proportionate to the number of non-redundant strains included in the database. *Homologous species groups* (i.e., species whose consensus sequences have more than 99% sequence similarity) were labeled with the same color. Strains that share 99% or higher sequence similarity in different species were linked by an edge.

To assign a species-level taxon to the OTUs, the representative sequence of each OTU was searched against our species consensus sequence database using BLAST search (Altschul et al., 1990). Similar to the *k*-nearest neighbor method, the species was determined by considering the most *k* homologous species. In this study, *k* was set to 1 among the sequences with >97% sequence similarity and an e-value of <1.0e-10. When no hit met the criteria, it was reported as unclassified. If the assigned species were from the *homologous species groups*, the query sequence was labeled as the name of the *homologous species group*.

Evaluation Using Simulated Datasets

To test the performance of our species-level profiling method, simulated datasets were generated using a set of bacteria reported as the constituents of the Human Microbiome Project (HMP) gut microbiome (**Supplementary Table 3**). The reported strains were downloaded from the NCBI RefSeq database, and the non-existing or updated strains were replaced with the latest strains of the same species. *Candida albicans* ATCC MY-2876 was not included since it is a fungus. Four simulated datasets were generated with the abundance models of uniform, linear, and power-law parameters with 1 and 2 (**Supplementary Table 3**). *Methanobrevibacter* and *Propionibacterium* (*Cutibacterium*) were excluded from the simulation since the 806R primer could not extract the region sequences from their genomes. The simulated datasets were analyzed using our species-level profiling method.

Regardless of the abundance model, the genus-level composition was almost perfectly profiled using our method (**Figure 4A**). Among the species in the simulated dataset, *Bacillus cereus*, *Bacteroides vulgatus, Clostridium beijerinckii*, *Escherichia coli*, *Lactobacillus gasseri*, *Listeria monocytogenes*, *Neisseria meningitidis*, *Pseudomonas aeruginosa*, *S. aureus*, *S. epidermidis*, and *Streptococcus pneumoniae* created the *homologous species groups* with other species. For instance, the V3–V4 region of *B. cereus* was identical to that of *Bacillus mobilis*. These species are technically indistinguishable in terms of their V3–V4 region. Similarly, the V3–V4 sequence of *S. pneumoniae* differs by only one nucleotide from that of *Streptococcus infantis*. Notably, our method based on the *homologous species groups* was able to accurately estimate the species-level composition in the simulated datasets (**Figure 4B**). Pearson's correlation coefficient values between the simulated and estimated bacterial composition were 0.9781 and 0.9790 for the genus- and species-level classification results. Therefore, our *homologous species groups* method could reasonably perform accurate species-level profiling.

Evaluation Using Mock Datasets

Six mock datasets consisting of 49 bacteria and 10 archaea (**Supplementary Table 4**) were downloaded from the EBI sequence repository[2] (Schirmer et al., 2015). The V3–V4 region was sequenced by Illumina MiSeq2 using the 341F forward primer and two kinds of reverse primers (806rcb and 805RA). The mock datasets were analyzed using our method (**Figure 5**). Since this mock data set provided a list of microbiome constituents without their relative abundance, we evaluated the

[2]http://www.ebi.ac.uk/ena/data/view/PRJEB6244

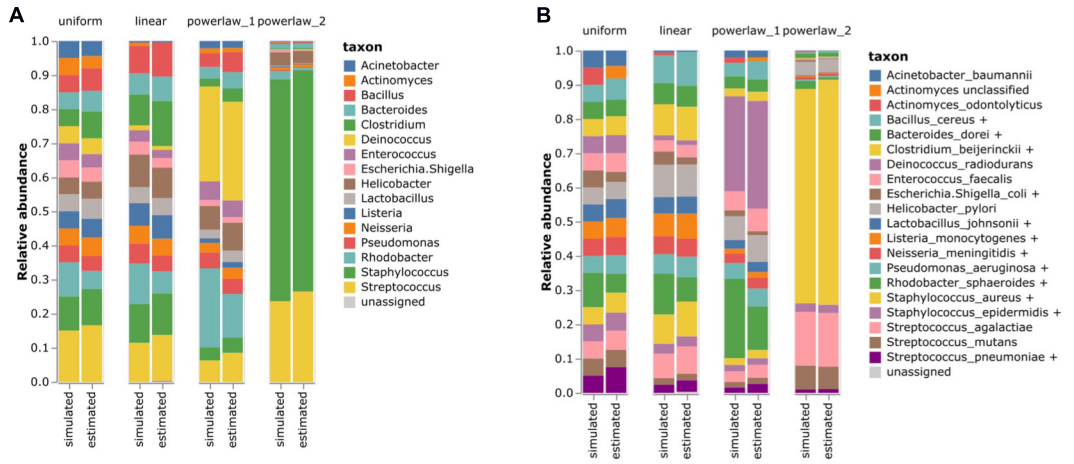


FIGURE 4 | Estimating the bacterial composition of the simulated datasets. **(A)** Genus level and **(B)** species level profiling. Simulated datasets with four different abundance models were analyzed using the proposed 16S rRNA classification pipeline. The names of the species that were contained in the simulated datasets were re-annotated according to the *homologous species groups* of the V3–V4 hypervariable region to compare the results.

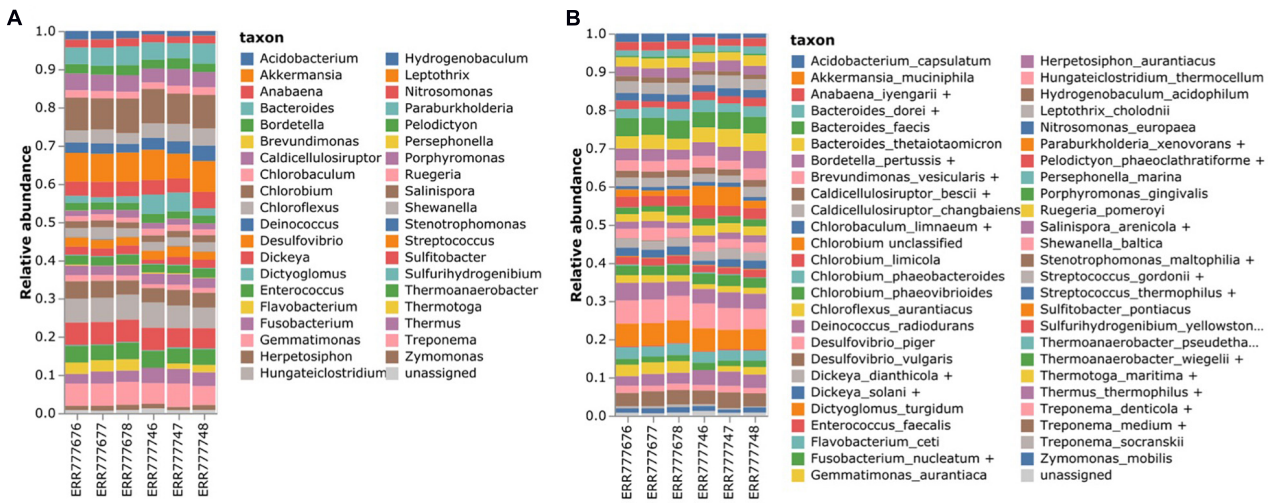


FIGURE 5 | Estimating the bacterial composition of the mock datasets. **(A)** Genus level and **(B)** species level profiling. Six mock samples were analyzed using the proposed 16S rRNA classification pipeline. The names of the species that were contained in the mock datasets were re-annotated according to the *homologous species groups* of the V3–V4 hypervariable region to compare the results.

results of our method by checking whether the specified genus and species were identified.

In total, 31 out of 38 genera were identified, accounting for an average of 90.7% of the microbiota population. In the case of *Burkholderia*, there were reads classified as *Paraburkholderia*. Moreover, *Anaerocellum* could not be identified owing to the lack of databases. On an average, 5.9% of the reads were misclassified as *Anabaena*, *Brevundimonas*, *Dickeya*, *Flavobacterium*, *Hungateiclostridium*, *Stenotrophomonas*, and *Streptococcus*. For the species-level classification, 31 out of 41 species were identified, of which 21 were assigned with specific species and 10 were assigned with *homologous species groups*. In total, 73.21% of the microbiota population on average was profiled at the species level. However, six species, namely *Anaerocellum thermophilum*, *Burkholderia xenovorans*, *Clostridium thermocellum*, and *Erwinia chrysanthemi*, could not be identified owing to the lack of databases.

A Case Study Using the Salivary Microbiome

In total, 90 salivary microbiome samples stratified by the oral hygiene index were downloaded from the DDBJ Sequence Read Archive (SRA) under the accession number DRA005425. A previous study reported that *Streptococcus* and *Veillonella* were the most abundant genera in these samples and that their proportions are associated with the hygiene index (Mashima et al., 2017). However, details regarding species-level

information were not provided. To profile the species-level composition, we re-analyzed the same salivary microbiome samples (**Figures 6A,B**). Notably, all of the *Streptococcus* and *Veillonella* OTUs were assigned to a species or *homologous species groups* (**Figures 6C,D**). With the exception of a few OTUs, most of the OTUs were assigned to the species groups. The *S. pneumoniae* group was identified as the most abundant species among all samples. Moreover, although the *S. pneumoniae* group was identified in both the good and poor hygiene groups, its abundance in the good hygiene group was more than twice that of the poor hygiene group.

In total, eight major *Streptococcus* OTUs were identified from the sample data by considering the size of the OTUs (number of reads in OTU >10,000): two OTUs with the *S. pneumoniae* group, three OTUs with the *Streptococcus gordonii* group, one OTU with the *Streptococcus sanguinis* group, one OTU with the *Streptococcus thermophilus* group, and one OTU with the *Streptococcus parasanguinis* group (**Figure 6C**). Two OTUs (OTU 221 and OTU 236) assigned to the *S. pneumoniae* group were equally similar to all *Streptococcus* species in the *S. pneumoniae* group, with the exception of *S. pneumoniae* as five species in the *S. pneumoniae* group have identical sequences in the 16S rRNA V3–V4 region, whereas *S. pneumoniae* differs by one nucleotide. Two OTUs (OTU 203 and OTU 411) assigned to the *S. gordonii* group also showed a similar pattern: they were equally similar to three species in the *S. gordonii* group. As shown in this case study, many OTUs were indistinguishable among the species in the species group but were distinguishable among the species group. Most of the *Veillonella* OTUs were assigned to the *Veillonella parvula*, *Veillonella dispar*, or *Veillonella atypica*

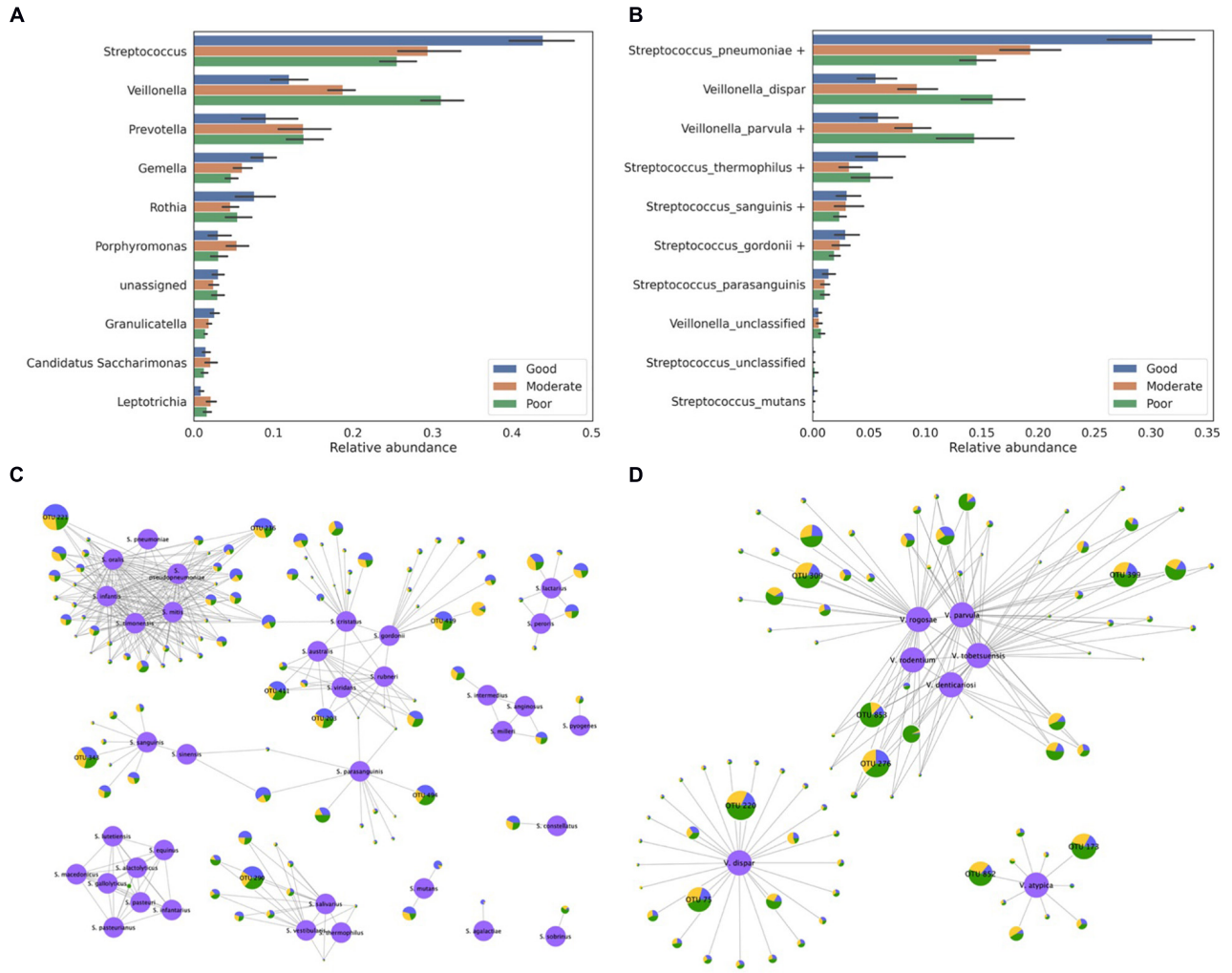


FIGURE 6 | Classification of the salivary microbiome samples obtained from three different hygiene index groups. **(A)** The relative abundance of the top 10 abundant genera and **(B)** the relative abundance of the species belonging to the *Streptococcus* and *Veillonella* genera are presented. The assignment results for **(C)** the *Streptococcus* and **(D)** the *Veillonella* genera are represented by the network. Reference sequences are colored in purple and labeled using their species name. Each node except the reference node represents one OTU, and the bacterial composition of samples from the three hygiene index groups is plotted as a pie-chart (blue for good, orange for moderate, and green for poor). All OTUs are represented as being proportional to their log-scaled size, and all OTUs with a size of 10,000 or more are labeled with their OTU number.

group (**Figure 6D**). In the *V. parvula* group, OTU 853 and OTU 276 were equally similar to multiple species in the group. These results might be inevitable when hypervariable regions are used at the species level. In addition, some novel species that are not stored in the 16S rRNA database but are equally similar to multiple known species could exist in the microbiome.

A Case Study Using the Gut Microbiome of Colon Cancer Patients

In total, 105 gut microbiome samples, consisting of 35 samples each from control, adenoma, and cancer patients, were downloaded from the SRA under the accession number SRP131074. *Bacteroides*, *Escherichia*, and *Prevotella* were reported as the most abundant genera in the previous study that analyzed these samples (Yang T. W. et al., 2019). Our results also showed that these three genera were the most abundant and in the same order (**Supplementary Figure 7**). Among the abundant genera, the abundance of *Megamonas*, *Pseudomonas*, *Morganella*, *Aeromonas*, *Megasphaera*, *Fusobacterium*, *Veillonella*, *Roseburia*, *Sutterella*, *Subdoligranulum*, and *Eubacterium* was found to differ by threefold between any two samples from the control, adenoma, and cancer groups (**Figure 7A**). Although most of the OTUs were assigned to a specific species without ambiguity, *Pseudomonas*, *Veillonella*, *Fusobacterium*, and *Aeromonas* OTUs were assigned to the *homologous species groups*. Notably, *Aeromonas* and *Fusobacterium* were the most abundant in the samples from the cancer group. For the *Aeromonas* OTUs, most of the dominant OTUs in the cancer group were assigned to either the *Aeromonas veronii* or *Aeromonas caviae* group (**Figures 7B,C**). Moreover, *Fusobacterium mortiferum*, *Fusobacterium necrophorum*, and *Fusobacterium nucleatum*

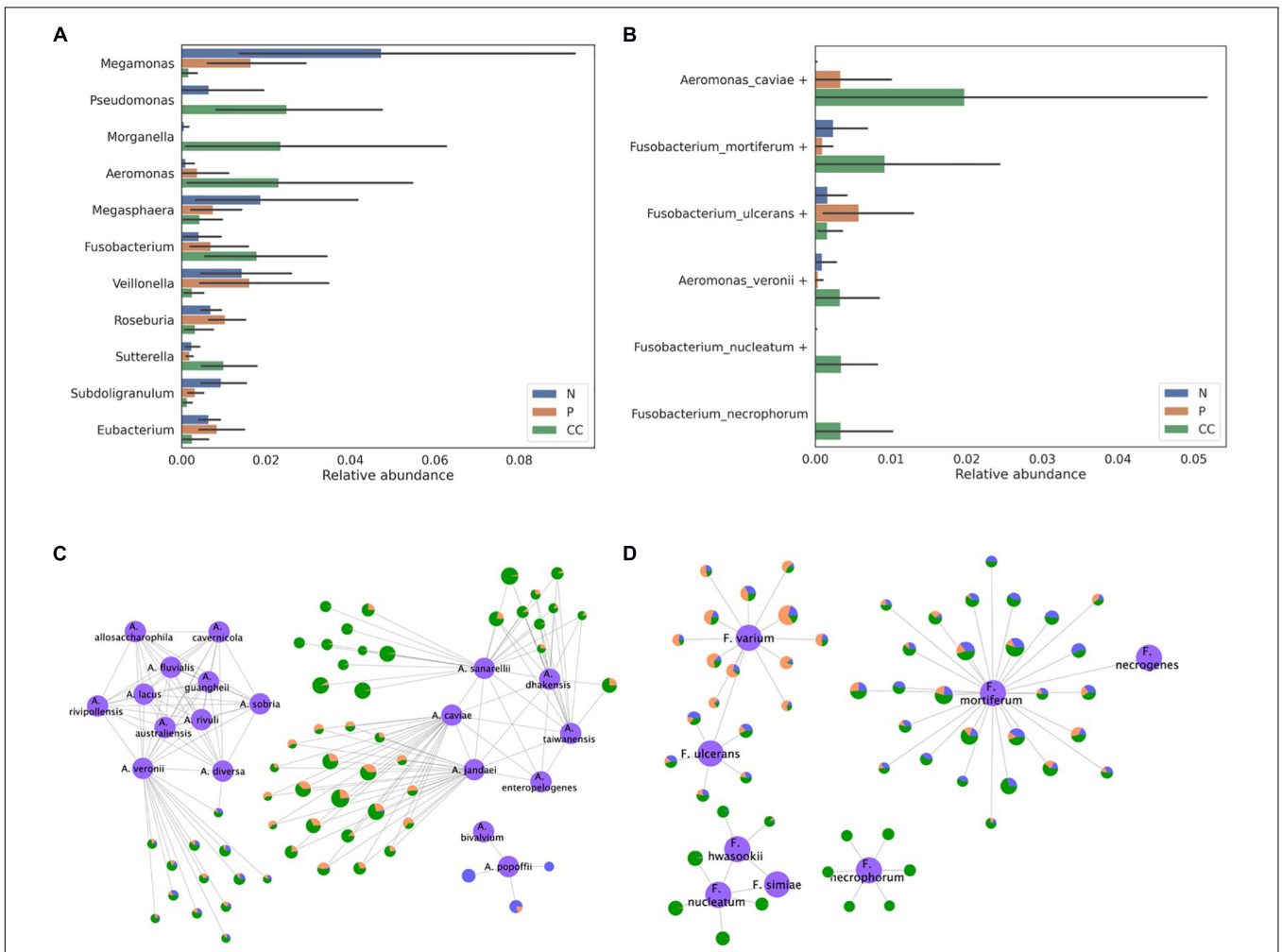


FIGURE 7 | Classification of the bacterial composition of gut microbiome samples obtained from the stool of control, adenoma, and cancer patients. **(A)** The relative abundance of the genera that were found to have an abundance more than three times that of the average relative abundance in at least one pair among the control, adenoma, and cancer groups and **(B)** the relative abundance of the species belonging to the *Aeromonas* and *Fusobacterium* genera are presented. The assignment results for the **(C)** *Aeromonas* and **(D)** *Fusobacterium* genera are represented by the network. The reference sequences are colored in purple and labeled using their species name. Each node except the reference node represents one out, and the bacterial composition of samples from each type of patient is plotted as a pie-chart (blue for control, orange for adenoma, and green for cancer). All OTUs are represented as being proportional to their log-scaled size.

were found to be abundant in samples from the cancer group, whereas *Fusobacterium ulcerans* was abundant in samples from the adenoma group (**Figures 7B,D**). Therefore, this indicates that our species-level profiling and network analysis based on *homologous species groups* could produce more specific and reliable information, which is higher resolution than the genus-level, to show differences in bacterial composition among patient groups.

CONCLUSION

In the microbiome studies, one of the important tasks is profiling of the bacterial composition, which helps understand the biological functions of the microbiome. The species-level taxonomic assignment is critical, but an optimal solution has not been available thus far since the 16S rRNA sequences are highly homologous between the species in the same genus in many cases. We combined all the sequences from the GreenGenes, SILVA, and NCBI databases to include species that exist exclusively in each database. Even in the evaluation of genus-level taxonomy classification, the classifier trained with the sequences combined showed the best accuracy in terms of precision and recall rates. For each species, we constructed a consensus sequence model and determined *homologous species groups*, which was used for the species-level taxonomy assignment. The evaluation using simulated datasets and mock datasets showed a high correlation with the real bacterial composition. When analyzing real gut microbiomes, our method successfully performed species-level taxonomic assignment and identified differential abundance between different phenotypic groups.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, re-annotated sequences and consensus sequences of each hypervariable region are available at https://sourceforge.net/projects/reannotated-16s-rrna-databases/files/.

AUTHOR CONTRIBUTIONS

H-JG designed and performed the data analysis and wrote the manuscript. MR designed and supervised the data analysis and wrote the manuscript. Both authors critically reviewed the manuscript and approved the final version.

FUNDING

This work was supported by the Bio & Medical Technology Development Program of the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT and Future Planning (2017M3A9F3041232 to MR), and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) [No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University)].

ACKNOWLEDGMENTS

We thank Ji-Hwan Ryu for the helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.570825/full#supplementary-material

REFERENCES

Allard, G., Ryan, F. J., Jeffery, I. B., and Claesson, M. J. (2015). SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics* 16:324. doi: 10.1186/s12859-015-0747-1

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., and Tyson, G. W. (2012). Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* 40:e94. doi: 10.1093/nar/gks251

Balvociute, M., and Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? *BMC Genomics* 18(Suppl. 2):114. doi: 10.1186/s12864-017-3501-4

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., et al. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* 185, 6220–6223. doi: 10.1128/jb.185.20.6220-6223.2003

Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., and Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLoS One* 15:e0228899. doi: 10.1371/journal.pone.0228899

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Cho, J. C., and Tiedje, J. M. (2001). Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Appl. Environ. Microbiol.* 67, 3677–3682. doi: 10.1128/AEM.67.8.3677-3682.2001

Chung, Y. W., Gwak, H. J., Moon, S., Rho, M., and Ryu, J. H. (2020). Functional dynamics of bacterial species in the mouse gut microbiome revealed by metagenomic and metatranscriptomic analyses. *PLoS One* 15:e0227886. doi: 10.1371/journal.pone.0227886

Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., et al. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* 68, 2386–2392. doi: 10.1099/ijsem.0.002809

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Cummings, P. J., Ahmed, R., Durocher, J. A., Jessen, A., Vardi, T., and Obom, K. M. (2013). Pyrosequencing for microbial identification and characterization. *J. Vis. Exp.* 78:e50405. doi: 10.3791/50405

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and

workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Douglas, H. C., and Gunter, S. E. (1946). The taxonomic position of *Corynebacterium acnes*. *J. Bacteriol.* 52, 15–23. doi: 10.1128/jb.52.1.15-23.1946

Dreno, B., Pecastaings, S., Corvec, S., Veraldi, S., Khammari, A., and Roques, C. (2018). Cutibacterium acnes (Propionibacterium acnes) and acne vulgaris: a brief look at the latest updates. *J. Eur. Acad. Dermatol. Venereol.* 32(Suppl. 2), 5–14. doi: 10.1111/jdv.15043

Edgar, R. (2018a). Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ* 6:e5030. doi: 10.7717/peerj.5030

Edgar, R. C. (2018b). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34, 2371–2375. doi: 10.1093/bioinformatics/bty113

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2015). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9, 968–979. doi: 10.1038/ismej.2014.195

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Res.* 40, D136–D143. doi: 10.1093/nar/gkr1178

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/MMBR.68.4.669-685.2004

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction. analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* 10:5029. doi: 10.1038/s41467-019-13036-1

Jumpstart Consortium Human Microbiome Project Data Generation Working Group (2012). Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS One* 7:e39315. doi: 10.1371/journal.pone.0039315

Khachatryan, L., de Leeuw, R. H., Kraakman, M. E. M., Pappas, N., Te Raa, M., Mei, H., et al. (2020). Taxonomic classification and abundance estimation using 16S and WGS-A comparison using controlled reference samples. *Forensic. Sci. Int. Genet.* 46:102257. doi: 10.1016/j.fsigen.2020.102257

Magoc, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10. doi: 10.14806/ej.17.1.200

Mashima, I., Theodorea, C. F., Thaweboon, B., Thaweboon, S., Scannapieco, F. A., and Nakazawa, F. (2017). Exploring the salivary microbiome of children stratified by the oral hygiene index. *PLoS One* 12:e0185274. doi: 10.1371/journal.pone.0185274

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453. doi: 10.1016/0022-2836(70)90057-4

Olson, S. A. (2002). EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief. Bioinform.* 3, 87–91. doi: 10.1093/bib/3.1.87

Peng, W., Huang, J., Yang, J., Zhang, Z., Yu, R., Fayyaz, S., et al. (2019). Integrated 16S rRNA sequencing, metagenomics, and metabolomics to characterize gut microbial composition, function, and fecal metabolic phenotype in Non-obese Type 2 Diabetic Goto-Kakizaki Rats. *Front. Microbiol.* 10:3141. doi: 10.3389/fmicb.2019.03141

Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A., and Versalovic, J. (2009). Metagenomic pyrosequencing and microbial identification. *Clin. Chem.* 55, 856–866. doi: 10.1373/clinchem.2008.107565

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341. doi: 10.1186/1471-2164-13-341

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Ravi, R. K., Walton, K., and Khosroheidari, M. (2018). MiSeq: a next generation sequencing platform for genomic analysis. *Methods Mol. Biol.* 1706, 223–232. doi: 10.1007/978-1-4939-7471-9_12

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43:e37. doi: 10.1093/nar/gku1341

Schloss, P. D., and Handelsman, J. (2003). Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* 14, 303–310. doi: 10.1016/s0958-1669(03)00067-3

Schloss, P. D., and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* 71, 1501–1506. doi: 10.1128/AEM.71.3.1501-1506.2005

Sessou, P., Keisam, S., Tuikhar, N., Gagara, M., Farougou, S., and Jeyaram, K. (2019). High-Throughput Illumina MiSeq amplicon sequencing of yeast communities associated with indigenous dairy products from republics of benin and niger. *Front. Microbiol.* 10:594. doi: 10.3389/fmicb.2019.00594

Stackebrandt, E. (2006). Taxonomic parameters revisited : tarnished gold standards. *Microbiol. Today* 8, 6–9. doi: 10.1016/0306-9192(84)90027-7

Sun, D. L., Jiang, X., Wu, Q. L., and Zhou, N. Y. (2013). Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. *Appl. Environ. Microbiol.* 79, 5962–5969. doi: 10.1128/AEM.01282-13

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wen, C., Wu, L., Qin, Y., Van Nostrand, J. D., Ning, D., Sun, B., et al. (2017). Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLoS One* 12:e0176716. doi: 10.1371/journal.pone.0176716

Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Braekel, J. V., et al. (2019). Targeting the 16s Rrna gene for bacterial identification in complex mixed samples: comparative evaluation of second (Illumina) and Third (Oxford Nanopore Technologies) generation sequencing technologies. *Int. J. Mol. Sci.* 21:298. doi: 10.3390/ijms21010298

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Yang, L., Haidar, G., Zia, H., Nettles, R., Qin, S., Wang, X., et al. (2019). Metagenomic identification of severe pneumonia pathogens in mechanically-ventilated patients: a feasibility and clinical validity study. *Respir. Res.* 20:265. doi: 10.1186/s12931-019-1218-4

Yang, T. W., Lee, W. H., Tu, S. J., Huang, W. C., Chen, H. M., Sun, T. H., et al. (2019). Enterotype-based analysis of gut microbiota along the conventional adenoma-carcinoma colorectal cancer pathway. *Sci. Rep.* 9:10923. doi: 10.1038/s41598-019-45588-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 19 November 2020
doi: 10.3389/fmicb.2020.596027

Human Gut Microbiome-Based Knowledgebase as a Biomarker Screening Tool to Improve the Predicted Probability for Colorectal Cancer

*Zhongkun Zhou[1], Shiqiang Ge[2], Yang Li[1], Wantong Ma[1], Yuheng Liu[1], Shujian Hu[1], Rentao Zhang[1], Yunhao Ma[1], Kangjia Du[1], Ashikujaman Syed[1] and Peng Chen[1]**

[1] School of Pharmacy, Lanzhou University, Lanzhou, China, [2] Department of Electronic Information Engineering, Lanzhou Vocational Technical College, Lanzhou, China

OPEN ACCESS

Edited by:
Hyun-Seob Song,
University of Nebraska-Lincoln,
United States

Reviewed by:
Minsuk Kim,
Mayo Clinic, United States
Anita Voigt,
The Jackson Laboratory for Genomic Medicine, United States

***Correspondence:**
Peng Chen
chenpeng@lzu.edu.cn

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: 18 August 2020
Accepted: 29 October 2020
Published: 19 November 2020

Citation:
Zhou Z, Ge S, Li Y, Ma W, Liu Y, Hu S, Zhang R, Ma Y, Du K, Syed A and Chen P (2020) Human Gut Microbiome-Based Knowledgebase as a Biomarker Screening Tool to Improve the Predicted Probability for Colorectal Cancer. Front. Microbiol. 11:596027. doi: 10.3389/fmicb.2020.596027

Colorectal cancer (CRC) is a common clinical malignancy globally ranked as the fourth leading cause of cancer mortality. Some microbes are known to contribute to adenoma-carcinoma transition and possess diagnostic potential. Advances in high-throughput sequencing technology and functional studies have provided significant insights into the landscape of the gut microbiome and the fundamental roles of its components in carcinogenesis. Integration of scattered knowledge is highly beneficial for future progress. In this study, literature review and information extraction were performed, with the aim of integrating the available data resources and facilitating comparative research. A knowledgebase of the human CRC microbiome was compiled to facilitate understanding of diagnosis, and the global signatures of CRC microbes, sample types, algorithms, differential microorganisms and various panels of markers plus their diagnostic performance were evaluated based on statistical and phylogenetic analyses. Additionally, prospects about current changelings and solution strategies were outlined for identifying future research directions. This type of data integration strategy presents an effective platform for inquiry and comparison of relevant information, providing a tool for further study about CRC-related microbes and exploration of factors promoting clinical transformation (available at: http://gsbios.com/index/experimental/dts_mben?id=1).

Keywords: biomarkers, colorectal cancer, database, diagnosis, microbiome

INTRODUCTION

Colorectal cancer (CRC) is a common malignancy worldwide accounting for about 1 in 10 cancer cases, with incidence and mortality rates of 6.1 and 9.2%, respectively (Bray et al., 2018). Various genetic and environmental factors contribute to CRC development from aberrant crypts to tumors. Overall, $\sim 3 \times 10^{13}$ bacteria colonize the human gut and abnormal microbiome composition has been shown to contribute to the initiation, progression and metastasis of CRC (Pitot, 1993; Qin et al., 2010; Wong et al., 2017c). In cases where patients are rapidly diagnosed and treated

with surgery at the early stages, survival exceeds 90%. However, the survival rate is significantly decreased to 13% in patients with advanced metastatic disease (Shah et al., 2018). The potential value of microorganisms in early diagnosis has attracted significant research attention over the last few decades.

The term "microbiome" refers to the entire habitat including microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes, and the surrounding environmental conditions (Marchesi and Ravel, 2015). These factors are altered along the adenoma-carcinoma sequence, reflected by changes in abundance. Some microbes produce genotoxic compounds and induce inflammation while others proliferate in the tumor-associated niche, designated "driver" and "passenger" bacteria, respectively (Tjalsma et al., 2012). Systematic analysis of microbial communities and identification of those with differential abundance as biomarkers presents an effective diagnostic strategy. Further advances, such as next-generation sequencing, have generated massive amounts of data on the CRC microbiome. Bioinformatics as well as machine learning methods additionally provide powerful tools to advance our understanding (Tabib et al., 2020). Metagenomics and 16S rRNA sequencing studies have revealed different abundance of some microbes between patients and healthy populations and effective combinations of microbial biomarkers could be applied for CRC diagnosis (Sze and Schloss, 2018; Thomas et al., 2019b). Upon combination of these strategies with the fecal immunochemical test (FIT), superior sensitivity and area under the receiver operating characteristic curve (AUC) were obtained relative to standalone FIT, which facilitated advanced adenoma detection (Wong et al., 2017a). Several microbes have been linked with CRC development, including *Fusobacterium nucleatum* (Fn), *Peptostreptococcus anaerobius* (Pa), *Parvimonas micra* (Pm), *Enterotoxigenic Bacteroides fragilis* (ETBF), *Peptostreptococcus stomatis* (Ps) and *Escherichia coli* (Yu et al., 2017a; Pleguezuelos-Manzano et al., 2020). Recently, the ratio of pathogenic bacteria to probiotic populations with decreased abundance in CRC patients was used in a diagnostic model based on their antagonistic effect (Guo et al., 2018). Metabolomics and metagenomics studies have shown that shifts in pathogenicity island genes, short-chain fatty acids (SCFA), amino acids, butyrate and bile acids occur at the early stages of CRC development. Some of these factors possess health-promoting and antineoplastic properties, such as maintenance of mucosal integrity and suppression of inflammation and carcinogenesis. Thus, the shift, particularly the decrease of these health-promoting factors, could contribute to the malignant outgrowth of the tumors (O'Keefe, 2016; Yachida et al., 2019). Subsequent mechanistic research further confirmed their involvement in CRC. For instance, Fn harbors the FadA virulence factor, which binds E-cadherin and activates Wnt/β-catenin and TLR4/MYD88 pathways to promote cancer initiation, proliferation and invasion (Rubinstein et al., 2013, 2019). Enterotoxigenic *Bacteroides fragilis*(ETBF) harbors the toxin BFT that causes inflammatory diarrhea, inflammation-related tumorigenesis and upregulation of spermine oxidase. Colibactin-producing *E. coli* alkylates DNA at adenine residues and induces double-stranded breaks, anaphase bridges and chromosome aberrations (Cuevas-Ramos et al., 2010; Goodwin et al., 2011; Chung et al., 2018; Pleguezuelos-Manzano et al., 2020). Based on these omics and experimental data, a theoretical foundation for clinical translation was proposed, which requires validation with more economical methods, such as quantitative PCR (qPCR), or integration with other indices, such as FIT, to obtain optimal benefits (Wong et al., 2017a). More novel biomarkers should emerge with further research progress. However, effective diagnostic panels remain to be established.

While several meta-analyses and reviews based on large-scale, cross-cohort studies have revealed robust associations between microbiome and diseases, developing solutions from the perspective of integration remains a considerable problem due to a number of reasons. First, among the published studies, feces is the most common sample type owing to the non-invasive nature and convenience of sample collection. Other non-invasive types of samples, such as oral swabs, offer an alternative but still need more studies (Flemer et al., 2018). Second, a number of studies were based on 16S rRNA sequencing while others involved metagenomics analyses, which may generate different taxonomic resolutions and involve distinct bioinformatics methods (Wirbel et al., 2019). Third, robustness among different countries or regions is another key contributory factor in microbiome composition, including genetic background, dietary habits and the environment. Fourth, optimal numbers of microbial markers recorded are significantly variable among studies (Duvallet et al., 2017). Fifth, specificity deserves further research attention, since only a few studies to date have included cases of other diseases. For example, *Helicobacter pylori* and human papillomavirus are specifically associated with gastric and cervical cancer types while other microbes, such as the order of *Clostridiales* (*Lachnospiraceae* and *Ruminococcaceae* families), are non-specifically associated with disease (Duvallet et al., 2017). In general, integration of different types of markers may obtain higher sensitivity, yet specificity will decrease. Therefore, biomarkers that are specific to CRC are of great importance. Finally, classification basis, algorithms, costs and standardization are also worth noting, but systematic integration of the data is lacking.

In this study, a knowledgebase of CRC-related microbes was established by reviewing the relevant literature and extracting key information. Next, a web-based platform using structured query language (SQL) was constructed and statistical analysis were performed that included three classifications and more than seven hundred records of microbial markers. By integrating the scattered data, our novel database could be used to perform inquiry and comparison across different models or databases, such as SILVA, VFDB and the Human Microbiome Oral Database (HOMD), thus contributing to the study of microbiome-based diagnosis of CRC.

MATERIALS AND METHODS

Database Construction

Literature was retrieved from PubMed during September 2019 and April 2020 based on the relevant search criteria.

Two keyword groups were used, the first being "colorectal cancer" and second comprising "16S rDNA," "metagenomics," "sequencing," "quantitative real-time PCR," "biomarker," "diagnosis," "screening," and "microbiome." Studies that used blood samples or focused on prognosis, genes, methylation, proteins, small molecule metabolites and liquid biopsy biomarkers were excluded. Following a comprehensive search of the literature and supplementary materials, the relevant data, including names of microbes, sensitivity, specificity, changes in abundance, functions of microbes, technology, algorithm, number of cases, sources and links, were collected. Furthermore, information of the taxonomy of microbial markers was collected from NCBI (Taxonomy) and added into the database. Ultimately, biomarkers were classified into three categories. Microbes that displayed statistical significance in both high-throughput sequencing/pyrosequencing and qPCR experiments were defined as "Class One," those confirmed with one of the above techniques as "Class Two," and combinations of different microbes for diagnosis as "Class Three." Notably, these candidates specifically refer to gut bacteria although the gut microbiome comprises bacteria, fungi, archaea, viruses and bacteriophages.

Data Query and Display

Integrated data were accessible through a web interface that indirectly generates MySQL queries. The interface supports query functions, such as "scientific name of the bacterium" and "taxonomy." Additionally, basic statistics and visualization were performed according to personalized requirements. Article links for verification or further research are provided for interested authors. The organizational framework is presented in **Figure 1**.

Construction of the Phylogenetic Tree and Statistical Analysis of CRC-Associated Microbes

16S rRNA sequences of all the species (all CRC-associated overabundant and depleted species) in the database were aligned using MEGA-X v10.1.8 software (Kumar et al., 2018). Phylogenetic tree was constructed using the following settings: maximum likelihood as the statistical method, 500 bootstrap replications, Kimura two-parameter as the substitution model and Near-Neighbor-Interchange as the ML Heuristic method. Finally, the tree was adjusted and visualized in Interactive Tree Of Life (iTOL)[1] (Letunic and Bork, 2019). Other statistical analyses were performed with OriginPro software (OriginLab Corporation, United States).

RESULTS AND DISCUSSION

Global Signature of CRC-Related Microbes

In our database, 17 species belonged to Class One (microbes with statistical importance verified using both high-throughput sequencing/pyrosequencing and qPCR), 219 species/clusters to Class Two (microbes confirmed via high-throughput sequencing/pyrosequencing or qPCR), including 11 phyla, 22 classes, 41 orders, 68 families and 117 genera (**Figure 2**), and 41 panels to Class Three (combinations of different microbes for diagnosis). Despite many microbes proposed for diagnosis and several confirmed conclusions, inconsistent results have been obtained by different research groups.

In healthy individuals, the most dominant phyla (over 90%) are *Firmicutes*, *Bacteroidetes, Proteobacteria* and *Verrucomicrobia* (Eckburg et al., 2005). Moreover, significant differences between healthy individuals and CRC patients are detected. Meanwhile, these differences of indices usually showed stepwise decreased or increased frequency from controls, to dysplasia to cancers, though some changes may not be statistically significant between healthy and adenoma groups. In addition to relative abundance, differences in other indices, such as alpha and beta diversity, have been identified. Feces of healthy controls generally contain microbial communities with higher diversity while tissue samples from CRC patients show greater alpha diversity. Earlier studies revealed greater microbial diversity in tumor samples compared with control and polyp samples, with a 75% higher estimated number of species than tissues from healthy sites (Mira-Pascual et al., 2015; Vogtmann et al., 2016), characterized by increased levels of opportunistic pathogens. Chao1 and Shannon indices are commonly used to estimate microbial richness and diversity. Decreased Shannon and Chao1 indices were recently reported in fecal samples collected from CRC patients (Yang et al., 2019). Similarly, in an azoxymethane (AOM) mouse model, the CRC group showed significantly lower bacterial richness and Shannon-Weaver's diversity index (Wong et al., 2017b). Other analyses revealed no significant differences in either richness or biodiversity, which could be attributable to the relatively small study cohorts (Wu et al., 2013; Youssef et al., 2018). However, differences at the taxonomic levels (family, genus and species) were universally observed. For instance, patients with CRC usually have increased abundance of operational taxonomic units (OTU) assigned as *Ruminococcus*, *Porphyromonas*, *Peptostreptococcus*, *Parvimonas*, and *Fusobacterium*, while healthy individuals possess more beneficial butyrate-producing bacteria, such as *Bifidobacterium* and *Clostridium butyricum* (Flemer et al., 2017; Sacks et al., 2018). The collective results clearly demonstrate differences in microbial populations between CRC and healthy groups.

Biomarker Identification for Diagnosis

Sample Types Used for Diagnosis

In studies on CRC-related microbes, fecal samples from CRC and adenoma patients and healthy volunteers were the most commonly used owing to the non-invasive nature and convenience of sample collection. Cancerous and adjacent non-cancerous normal tissues represent another type of sample that can effectively reveal the overall structure of microbiota in the tumor microenvironment but are unsuitable for early diagnosis (Gao et al., 2015). The microbial diversity in fecal samples is twice as high as that in tissue samples (Mira-Pascual et al., 2015). Oral

[1] https://itol.embl.de/

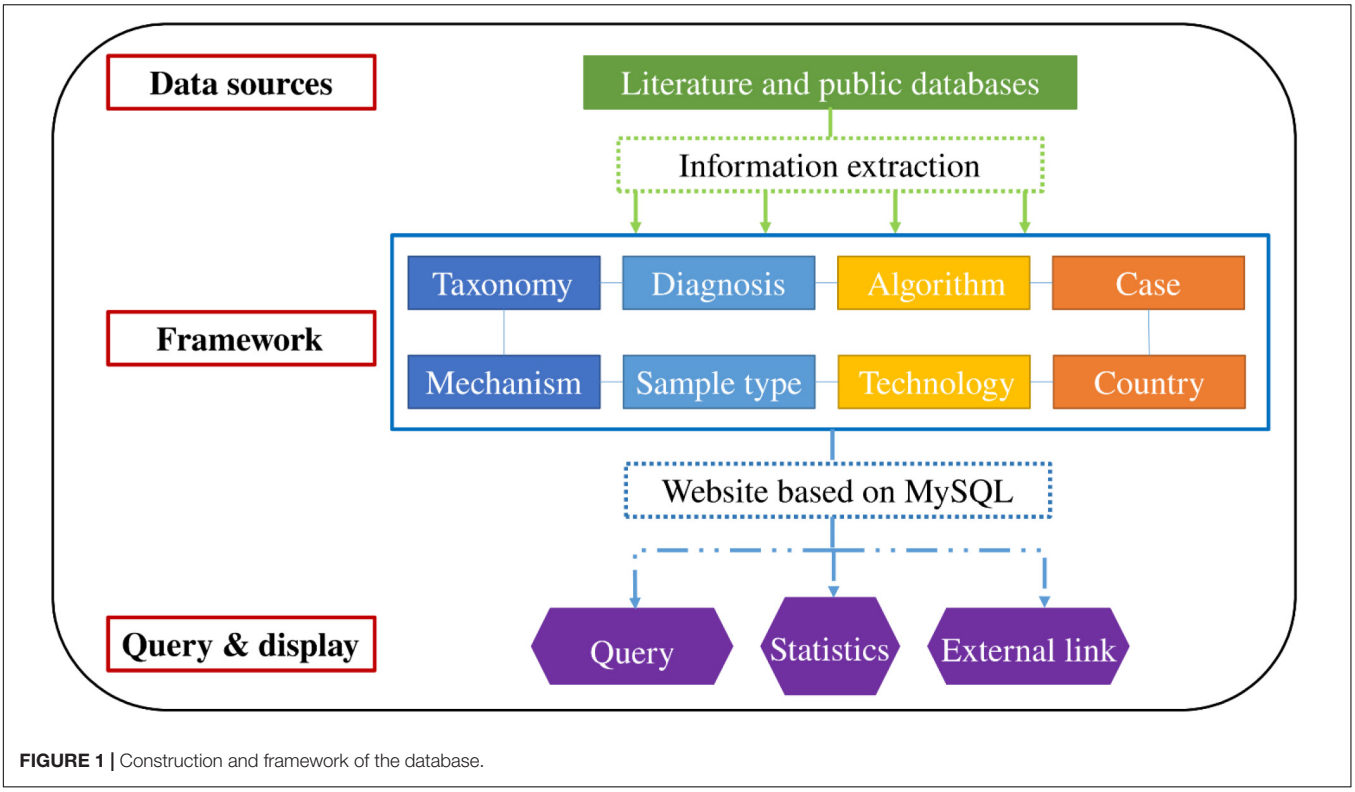


FIGURE 1 | Construction and framework of the database.

swabs represent another novel sample type. Previously identified biomarkers, such as *Fusobacterium nucleatum* and *Parvimonas micra*, are oral microbes. An earlier investigation profiled the oral microbiome as an alternative screening method for CRC (Flemer et al., 2018). Interestingly, a retrospective study on data obtained from adult patients diagnosed with bacteremia and subsequently CRC reported association with *Bacteroides fragilis*, *Streptococcus gallolyticus* and other intestinal microbes, thus providing a new perspective for clinicians (Kwong et al., 2018). Recently, (Poore et al., 2020) reported that predictions based on microbial DNA in blood could discriminate CRC from healthy, cancer-free individuals. However, blood samples were not included in this database due to the requirement for further exploration.

Diagnostic Techniques

This database involves five technical protocols, specifically, denaturing gradient gel electrophoresis (DGGE), qPCR, pyrosequencing, 16S rRNA sequencing and metagenomics sequencing, which have various advantages and disadvantages. Initially, the culture-dependent method was used to analyze CRC microbes as early as the 1960s, which led to significant underestimation of microbial diversity (Wong and Yu, 2019). Recently, a library containing 7,758 human gut bacterial isolates was constructed. Although culture-based methodologies provide access to data that both overlap and complement sequencing surveys, yet these protocols were both labor- and time-consuming compared with culture-independent methods (Poyet et al., 2019). Molecular analysis technology has developed from DGGE and qPCR to high-throughput sequencing over the years. While the efficiency of analysis was improved by DGGE and qPCR, limitations of low throughput remained unresolved. In 2005, the introduction of next-generation sequencing (NGS) facilitated massive parallel, low-cost and rapid sequencing. 16S rRNA and metagenomics sequencing have further improved efficiency and are widely employed at present. The former procedure is based on the 16S rRNA gene amplicon and facilitates taxonomic and phylogenetic analyses. While the cost-effective feature enables its universal application, several limitations exist: (1) amplicon sequencing of 16S rRNA gene via PCR may miss OTU/taxa detection due to various biases associated with PCR, (2) possible overestimation of community diversity or species abundance, and (3) lack of ability to directly analyze the biological functions of associated taxa (Xia et al., 2018). Recently, potentially unbiased shotgun metagenomics analyses have been conducted, which provide higher taxonomic resolution, gene function and comparative analyses at a decreased cost (Wirbel et al., 2019). However, in terms of clinical transformation, the qPCR-based method is more economical and rapid.

Algorithms Used for Diagnosis

Algorithms include the processes of classification, biomarker identification and model prediction. The classification approaches comprise OTU-based, metagenomics linkage group (MLG)-based, integrated microbial genome (IMG)-based and co-abundant gene group (CAG)-based methods. The model prediction algorithms include random forest (RF), support

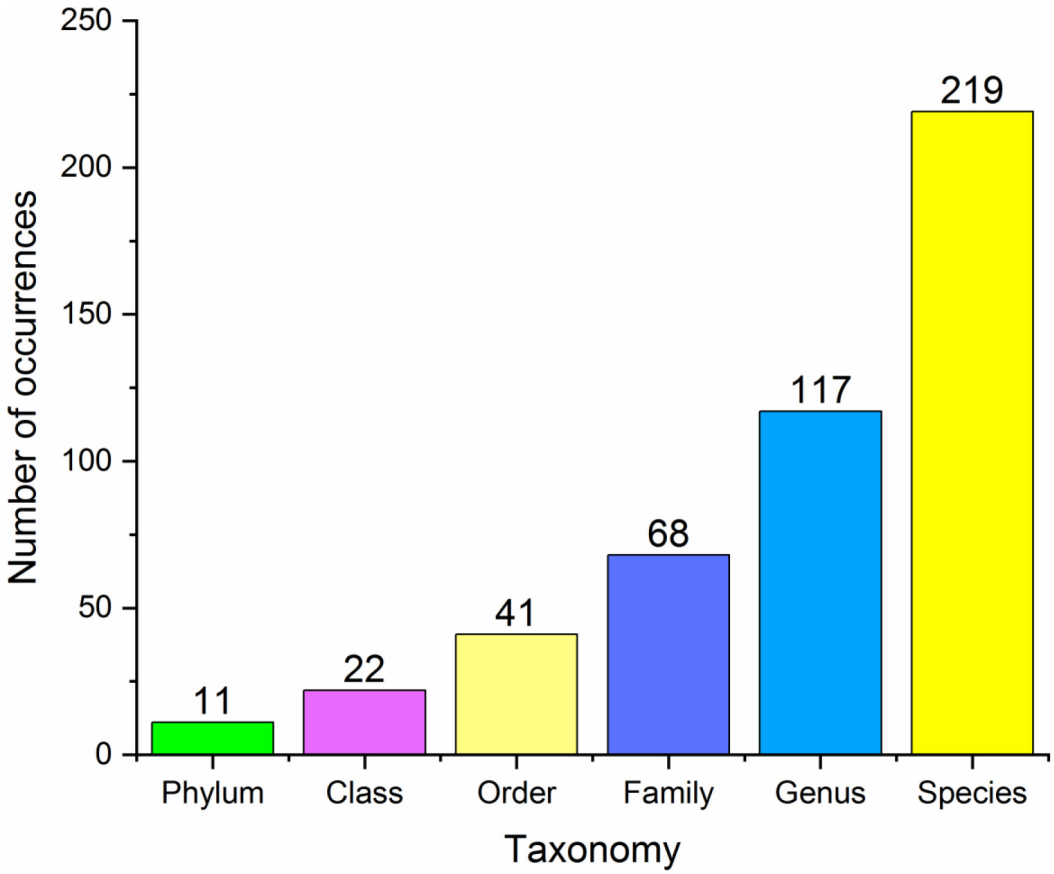


FIGURE 2 | Basic statistics at different taxonomy levels of all the microbial markers in the database.

vector machine (SVM), logistic regression (LR) and leave-one-dataset-out (LODO) analyses, among which random forest is the most widely used algorithm. For the biomarker identification process, relative abundance and Linear discriminant analysis Effect Size (LEfSe) methods are the most commonly used.

Random forest provides a measure of variable importance and out-of-bag (OOB) error when building a tree, making it suitable for prediction analysis. A recent meta-analysis employed the random forest classifier to determine accurate predictive models using a minimal microbial signature. The data showed that using 16 species, cross-validation of AUC $>$ 0.80 was achieved for the majority of datasets (Thomas et al., 2019a). SVM is advantageous for classifying small data volumes and achieved an overall AUC of 0.80 for the combined population (Dai et al., 2018). Recent studies have examined different machine leaning classifiers, including RF, Bayesian network, SVM, k-Nearest neighbor and general regression neural networks (Arabameri et al., 2020). LR, applied by most studies, is used to predict binary outcome from a set of numeric variables and aims to identify the most significant features (Wong et al., 2017a). Phylotype-based and OTU-based methods are the main approaches for sequence identification, with the latter being most widely used. However, the OTU-based method has a number of limitations, such as a computationally intensive protocol and larger memory requirement (Schloss and Westcott, 2011). Other methods have been developed to overcome these drawbacks. For instance, CAGs have been proposed to mitigate the ultrahigh dimensionality challenge of gene-level metagenomics (Minot and Willis, 2019). In addition, CAG-based clusters could be used to determine CRC-associated microbe profiles (Flemer et al., 2017). Taking the collective factors (such as data quantity, number of cohorts and risk factors) into consideration, appropriate approaches and classifiers should be adopted.

Overview of Current Biomarkers for Diagnosis

More than 200 species belonged to the Class Two microbe group (confirmed using either high-throughput sequencing/pyrosequencing or qPCR), among which only 17 were verified as statistically significant with both high-throughput sequencing/pyrosequencing and qPCR (Class One). Fn is a known opportunistic pathogen showing increased abundance in feces of CRC patients with a sensitivity range of 69.2–82.9%, specificity of 52.8–90.8% and AUC of 0.675–0.875. Combined with FIT or fecal occult blood test (FOBT), sensitivity, specificity and AUC values reached 92.3, 94.4% and 0.95, respectively. Recently, a number of novel markers have been shown to perform well in CRC diagnosis. Pa was increased in four different cohorts and induced carcinogenesis in mice via a

PCWBR2-integrin α2/β1-PI3K–Akt–NF-κB signaling axis with a sensitivity of 79.8% and specificity of 98% in combination with FIT (Yu et al., 2017a; Long et al., 2019). *Lachnoclostridium sp.* (designated m3) sharing 97% (1883/1935) DNA sequence similarity with *Lachnoclostridium sp.* YL32 was significantly enriched in adenoma. m3 showed specificity of 78.5% and sensitivity of 48.3% for adenoma and 62.1% for CRC. However, its role in tumorigenesis warrants further research (Liang et al., 2019). The other 15 biomarkers are presented in **Table 1** (4 were decreased and 11 were enriched in patients).

With regard to Class Two microbes, basic statistics are shown in **Figure 3** and phylogenetic tree in **Figure 4**. The majority of enriched microbes were classified into *Fusobacteriaceae*, *Peptoniphilaceae*, *Lachnospiraceae*, *Porphyromonadaceae*, *Peptostreptococcaceae*, *Bacteroidaceae*, *Prevotellaceae, Ruminococcaceae, Streptococcaceae*, and *Bacillales incertae sedis* at the family level (**Figure 3A**). Among the group of decreased microbes, most were classified into *Lachnospiraceae*, *Ruminococcaceae*, *Bacteroidaceae, Streptococcaceae, Bifidobacteriaceae*, and *Eubacteriaceae* (**Figure 3B**). In the Venn diagram, only a small overlap of increased and decreased microbes was observed, supporting the reliability of most microbial markers despite some inconsistencies (**Figure 3C**). At the species level, phylogenetic tree showed details of current CRC-related biomarkers as well as their evolutionary relationships. Additionally, species belonging to oral microbes were marked with stars.

The functions of gut microbes include fermenting complex carbohydrates to produce large amounts of metabolites, maintaining epithelial homeostasis, serving as an endocrine organ and participating in the development, maturation and differentiation of the immune system of the host (Villéger et al., 2018; Rastelli et al., 2019). In a sense, intestinal metabolites directly affect the occurrence of CRC and not intestinal flora. The majority of nutrients from food are absorbed in the small intestine with protein residues and complex nutrients, such as fiber moving to the colon, and consequently metabolized by the microbial populations (O'Keefe, 2016). Therefore, from the perspective of microbial function, the majority are associated with protein fermentation, bile acid biotransformation, decomposition of polysaccharides and polyphenols and energy metabolism. For example, *Faecalibacterium prausnitzii* (Fp), *Bifidobacterium* (Bb), *Roseburia* spp. (Rb), *Eubacterium rectale* (EUB), *Clostridium butyicum* (Cb), *Lactobacillus* spp. (Lc), *Akkermansia muciniphila* (Akk), *Ruminococcus*, and *Lachnospiraceae* were found to be more abundant in healthy controls compared with CRC patients. Fp is a butyrate producer decreased in Crohn's disease (CD) patients, whose metabolites exert anti-inflammatory effects via blocking NF-κB activation and IL-8 production (Sokol et al., 2008). Bb and Lc are used as probiotics for human consumption and benefit the gut through inducing cancer cell apoptosis, inhibiting cell proliferation, modulating host immunity and inactivating carcinogenic toxins (Wong and Yu, 2019). An earlier study reported that determination of Fn/Bb and Fn/Fp ratios could improve diagnostic performance for CRC based on their antagonistic effect (Rezasoltani et al., 2018). Both Rb and EUB are butyrate-producing *Firmicutes* and metabolize dietary fibers to provide energy sources and achieve anti-inflammatory effects (Paramsothy et al., 2019). Their capabilities as a non-invasive tool were additionally evaluated but not included in the final model (Malagón et al., 2019). More recently, the utility of other widely recognized markers, including Fn, colibactin-producing *E. coli* and ETBF, in diagnosis of CRC has been systematically analyzed (Chung et al., 2018; Malagón et al., 2019; Wu et al., 2019; Pleguezuelos-Manzano et al., 2020). However, several

TABLE 1 | Diagnostic performance of Class One microbials.

Name	Sensitivity%	Specificity%	AUC	Algorithm	Sample	Case	Region	References
Fn	73.1	90.8	0.860	Relative Abundance	Feces	490	China	Wong et al., 2017a
Pa	56.7	86.3	0.720	Logistic regression	Feces	390	China	Wong et al., 2017a
Pm	45.2	97.1	0.730	Logistic regression	Feces	390	China	Wong et al., 2017a
Gm	39.0	76.0	0.622	Relative Abundance	Feces	333	Spain	Malagón et al., 2019
Ps	53.0	76.0	0.710	Relative Abundance	Feces	333	Spain	Malagón et al., 2019
Bf	33.0	0.76	0.571	Relative Abundance	Feces	333	Spain	Malagón et al., 2019
pks	56.4	82.0	NA	Relative Abundance	Feces	238	Sweden	Eklöf et al., 2017
Fp	81.8	62.6	0.741	Abundance Rate	Feces	549	China	Rezasoltani et al., 2018
Bb	90.4	76.4	0.870	Abundance Rate	Feces	549	China	Rezasoltani et al., 2018
Cs	73.3	66.1	0.736	logistic regression	Feces	781	China	Xie et al., 2017
Ap	NA	NA	NA	Relative abundance	Feces	146	Meta	Yachida et al., 2019
Gl	NA	NA	NA	Relative abundance	Mucosa	207	China	Nakatsu et al., 2015
m3	62.1	79.0	0.741	Relative Abundance	Feces	1012	China	Liang et al., 2019
Bd	NA	NA	NA	Relative abundance	Feces	179	French	Sobhani et al., 2011
afaC	NA	NA	NA	Relative abundance	Tissue	55	South Africa	Viljoen et al., 2015
Akk	NA	NA	NA	Relative abundance	Feces	112	China	Wang et al., 2020
Cb	NA	NA	0.930	Random forest	Feces	60	China	Yang et al., 2020

NA, non-available; Meta, meta-analysis; Gemella morbillorum (Gm); Bacteroides fragilis (Bf); pks + clbA + Escherichia coli (pks); Clostridium symbiosum (Cs); Atopobium parvulum (Ap); Granulicatella (Gl); Bacteroides (Bd); afaC-positive E. coli (afaC).

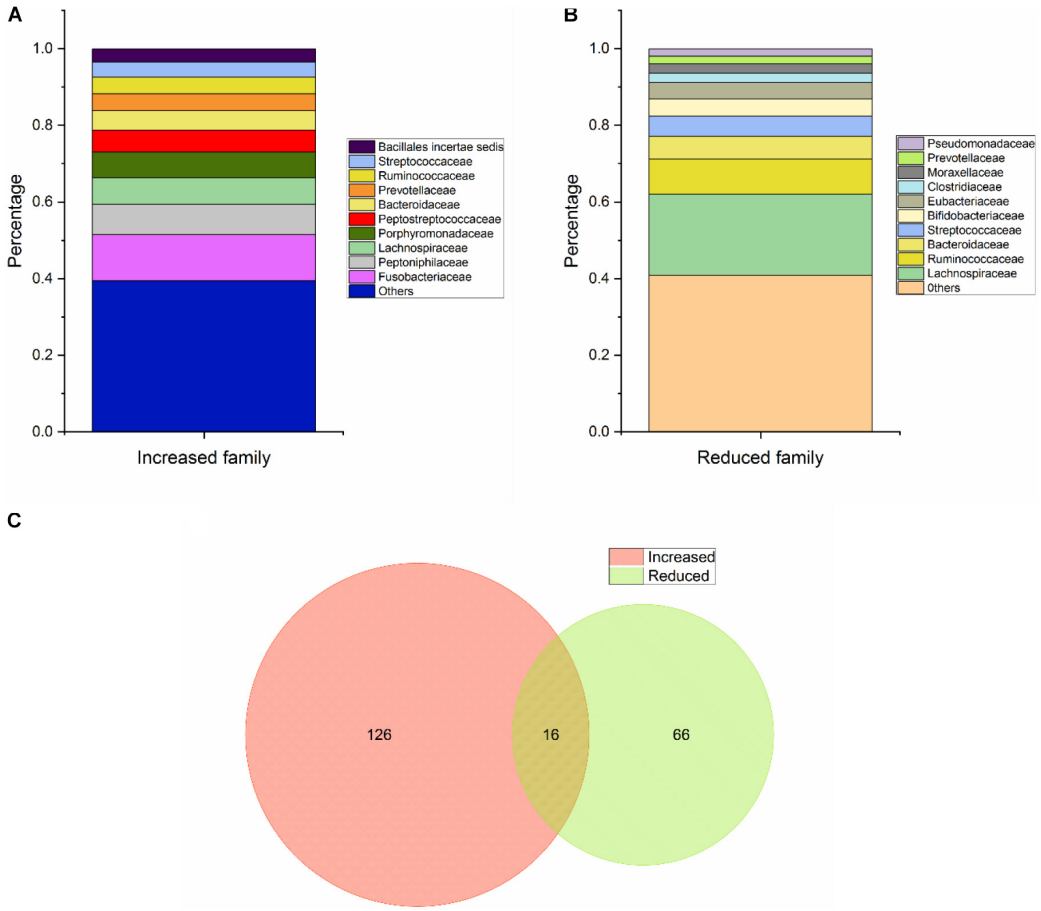


FIGURE 3 | Basic statistical analysis of Class Two microbes (shown to be significant via high-throughput sequencing/pyrosequencing or qPCR) in the database. **(A)** Statistical analysis of the top 10 increased microbes at the family level. **(B)** Statistical analysis of the top 10 reduced microbes at the family level. **(C)** Venn diagram of all CRC-associated microbes at the species level.

issues require further clarification. Although the pathogenesis and benefits of ETBF and Bb have been validated, inconsistencies exist among different samples. ETBF was shown to be increased in tumor tissues and form a biofilm in the gut. However, this pathogenic bacterium displayed no significant differences in abundance in patient fecal samples and was not detectable using qPCR targeting the toxin-producing gene, making it difficult to discriminate between patients and healthy controls (Zackular et al., 2014; Kosumi et al., 2018; Sze and Schloss, 2018; Malagón et al., 2019; Saffarian et al., 2019). Finally, *Lachnospiraceae* and *Ruminococcaceae* families were associated with multiple diseases (known as non-specific responders), which inspired us to obtain non-gastrointestinal cancer samples for future experimental design (Duvallet et al., 2017; Rezasoltani et al., 2018).

Diagnostic Strategy and Performance

Combinations of Different Microbial Markers

Class Three (combinations of different microbes for diagnosis) included 41 panels verified using various methods (**Table 2**). The combinations ranged from two species to 63 OTUs, with AUC ranging from 0.531 to 0.998. Twelve panels were based on qPCR, whose algorithms usually link with logistic regression or relative abundance. Meanwhile, 16 panels and 12 combinations were based on 16S rRNA and metagenomics sequencing data, predominantly using the random forest-based model. Based on AUC, qPCR-based models could achieve comparable outcomes to the two other technologies with limited biomarkers (usually no more five species). Nevertheless, 16S rRNA and metagenomics-based models show performance advantages at the cost of the number of markers (more than 10 OTUs on average). In the random forest and Minimum Redundancy Maximum Relevance (mRMR) models, both OOB and error rate parameters demonstrated that panels comprising ∼16–20 biomarkers achieved the best prediction accuracy (Flemer et al., 2018; Wirbel et al., 2019).

Combination of microbes may be operative, rather than representing a strain that is increased or decreased in the intestine (Tilg et al., 2018). In addition, prediction models from single dataset may lead to reduced accuracy and be sensitive to both technique and heterogeneity (Thomas et al., 2019a). An earlier study identified 63 OTUs (29 from oral swabs and 34 from fecal samples) to predict CRC. While the final AUC value

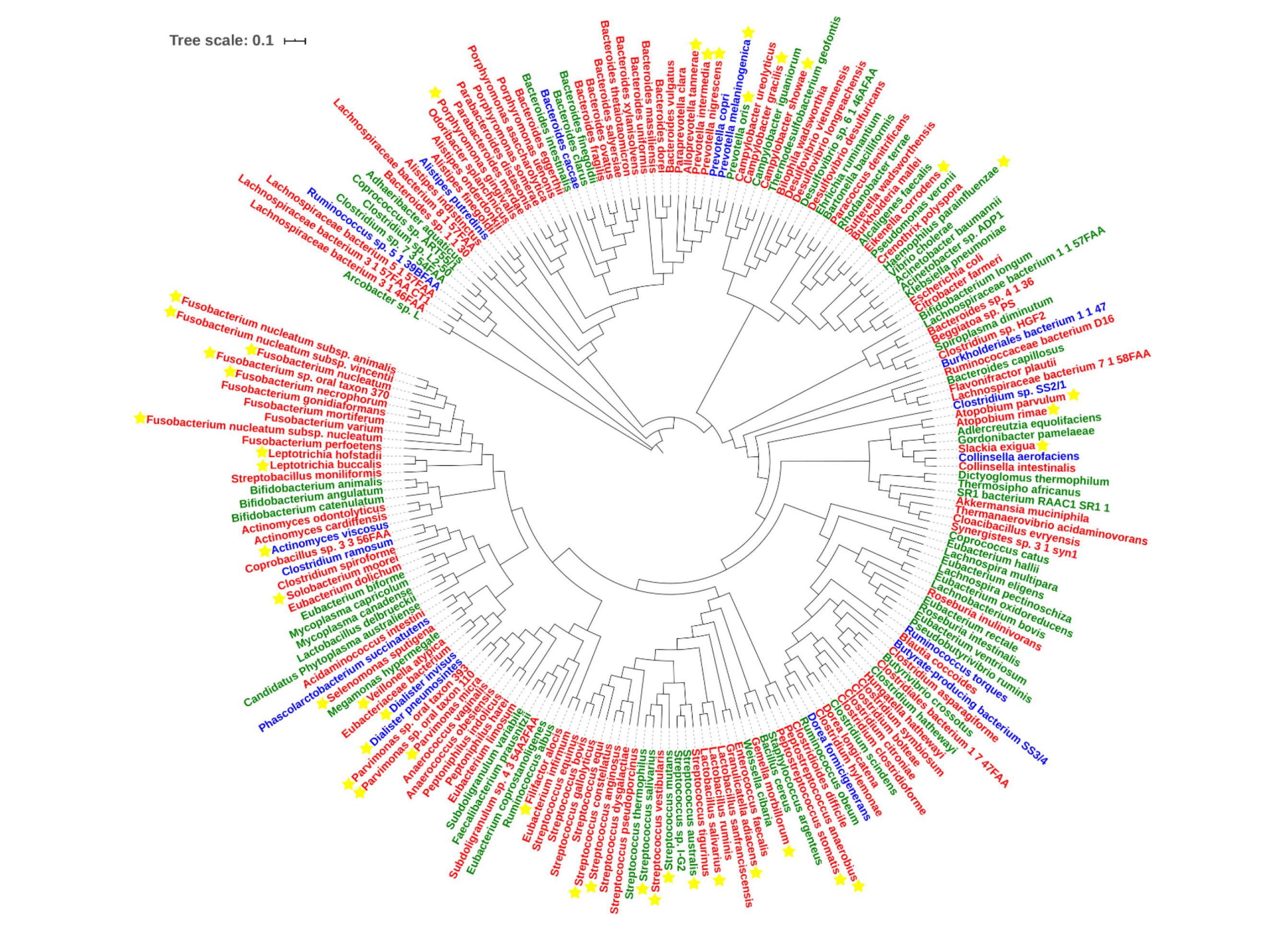


FIGURE 4 | Phylogenetic tree of all CRC-related microbes in the database. Species marked in red and green refer to the increased and decreased microbes, and species marked in blue refer to the microbes that show up in both increased and decreased groups. Species marked with yellow stars refer to oral microbes according to HMOD (16S rRNA sequences of m7 and *Sulfurovum* sp. SCGC AAA036-O23 are not available, which also belong to the increased group).

was up to 0.98, its application in clinical examination remains a challenge (Flemer et al., 2018). Several other researchers used more than 30 OTUs/phylotypes/MLGs to construct a random forest classifier and obtained AUC values > 0.80 (Nakatsu et al., 2015; Baxter et al., 2016a; Yu et al., 2017a). Previous studies suggest that the *Firmicutes*/*Bacteroidetes* ratio responds to health and disease states, such as obesity and CRC (Ley et al., 2006; Saffarian et al., 2019). Interactions between bacteria provide an ecological perspective for screening, and increase in pathogenic bacteria is always accompanied by decrease in beneficial microbes (Dai et al., 2018). Some researchers observed an association of the group of *Bacteroides* and *Prevotella* with elevated IL17-producing cells in colon cancer and demonstrated that supernatant from Fn inhibited the bactericidal activities of Fp and Bb (Sobhani et al., 2011; Guo et al., 2018). Furthermore, beneficial microbes can contribute to several intestinal functions and protect the organ from pathogenic microorganisms, and the "pathogenic bacteria:probiotics" ratio generates a better effect than single organism model (Eslami et al., 2019; Malagón et al., 2019; Yang et al., 2020). Thus, the complementary effects between enriched and reduced microbes should be highlighted for further investigation. Clearly, combinations of different microbial markers exhibit better predictive performance than single markers.

Integration With FIT

In the database, FIT was also presented when available. FIT has been extensively tested and recommended by National Comprehensive Cancer Network guidelines. The method involves direct detection of globin rather than heme, and shows greater sensitivity than the highly sensitive guaiac fecal occult blood test. Retrospective analysis showed that replacing 3-year colonoscopy surveillance with annual FIT could reduce the requirement for colonoscopy and provide economic benefits. However, sensitivity was relatively low for advanced neoplasms, ranging from 21.8 to 46.3% at the preset thresholds (Gies et al., 2018; Cross et al., 2019a). Combining microbe analysis with FIT could enhance the detection of advanced precancerous lesions, as validated in numerous experiments. Taking results from Class One and Three as representative cases, combined quantitation

TABLE 2 | Different panels for CRC screening.

Name	Sensitivity%	Specificity%	AUC	Technique	Algorithm	Sample	Case	Region	References
Fn, Pa, Pm	89.4	93.0	0.950	qPCR	LR	Feces	390	China	Wong et al., 2017a
Ps/EUB, Bf/EUB, Bt/EUB	80.0	90.0	0.837	qPCR	LR	Feces	333	Spain	Malagón et al., 2019
pks, Fn	89.7	61.0	NA	qPCR	DA	Feces	238	Sweden	Eklöf et al., 2017
Fn/Fp	95.0	71.3	0.914	qPCR	AR	Feces	549	China	Rezasoltani et al., 2018
Fn/Bb	84.6	92.3	0.911	qPCR	AR	Feces	549	China	Rezasoltani et al., 2018
Fn/Fp, Fn/Bb	80.8	85.6	0.910	qPCR	AR	Feces	549	China	Rezasoltani et al., 2018
Fn, Fp, Bb	92.5	83.5	0.943	qPCR	AR	Feces	549	China	Rezasoltani et al., 2018
5 OTUs	90.0	80.0	0.896	16SrDNA	LR	Feces	90	America	Zackular et al., 2014
6 OTUs	90.0	83.0	0.922	16SrDNA	LR	Feces	90	America	Zackular et al., 2014
22OTUs	81.2	97.1	0.673	16SrDNA	RF	Feces	490	Canada, United States	Baxter et al., 2016b
34 OTUs	51.7	97.1	0.847	16SrDNA	RF	Feces	490	Canada, United States	Baxter et al., 2016b
23 OTUs	70.0	92.8	0.829	16SrDNA	RF	Feces	490	Canada, United States	Baxter et al., 2016b
16 OTUs	53.0	96.0	0.900	16SrDNA	RF	Oral swabs	60	Ireland	Flemer et al., 2018
28 OTUs (16 oral swabs, 12 feces)	74.0	94.0	0.940	16SrDNA	RF	Feces and oral swabs	60	Ireland	Flemer et al., 2018
63 OTUs (29 oral swabs, 34 feces)	88.0	94.0	0.980	16SrDNA	RF	Feces and oral swabs	60	Ireland	Flemer et al., 2018
22 OTUs	58.0	92.0	0.840	Metagenomics	LR	Feces	156	France, Germany	Zeller et al., 2014
7 OTUs	87.0	83.7	0.886	Metagenomics	RF	Feces	128	China	Yu et al., 2017a
15 MLGs	NA	NA	0.983	Metagenomics	RF	Feces	96	Austria	Feng et al., 2015
16 OTUs	NA	NA	0.860	Metagenomics	RF	Feces	969	Meta	Thomas et al., 2019a
17 OTUs	60.1	84.8	0.804	Metagenomics	RF	Feces	424	Meta	Shah et al., 2018
30 OTUs	NA	NA	0.830	Metagenomics	RF	Feces	208	Meta	Yachida et al., 2019
8 taxa	NA	NA	0.750	16SrDNA	RF	Feces	492	Meta	Yachida et al., 2019
12 genus	NA	NA	0.846	16SrDNA	RF	Feces	1674	Meta	Sze and Schloss, 2018
18 OTUs	NA	NA	0.831	16SrDNA	RF	Feces	404	Canada, United States	Baxter et al., 2016a
32 OTUs	NA	NA	0.853	16SrDNA	RF	Feces	404	Canada, United States	Baxter et al., 2016a
41 OTUs	NA	NA	0.686	16SrDNA	RF	Feces	404	Canada, United States	Baxter et al., 2016a
12 phylotypes	NA	NA	0.831	16SrDNA	LEfSe	Mucosa	160	China	Nakatsu et al., 2015
18 OTUs	NA	NA	0.871	16SrDNA	RF	Mucosa	160	China	Nakatsu et al., 2015
38 phylotypes	NA	NA	0.846	16SrDNA	Dirichlet MM	Mucosa	160	China	Nakatsu et al., 2015
m3, Fn, Ch, Bc	85.2	80.2	0.907	qPCR	LR	Feces	1012	China	Liang et al., 2019
m3, Fn	NA	NA	0.891	qPCR	LR	Feces	1012	China	Liang et al., 2019
Fn, Ch, m7, Bc	92.8	79.8	0.886	qPCR	SLC	Feces	370	China	Liang et al., 2017
Fn, Ch, m7, Bc, Ri	74.3	88.9	0.843	qPCR	LR	Feces	128	China	Liang et al., 2017
17 IMG species	NA	NA	0.860	Metagenomics	IMG	Feces	128	China	Liang et al., 2017
7 species-level mOTUs	NA	NA	0.890	Metagenomics	mOTUs	Feces	128	China	Liang et al., 2017
27 MLG	NA	NA	0.960	Metagenomics	MLG	Feces	128	China	Liang et al., 2017
Fn, Pa, Pm (4 genes)	NA	NA	0.770	Metagenomics	CRC index	Feces	96	China, Denmark, Austrian, French	Yu et al., 2017a
22 genes	NA	NA	0.998	Metagenomics	RF	Feces	107	China	Yang et al., 2020
Cb, Cs	NA	NA	0.935	qPCR	RF	Feces	60	China	Yang et al., 2020
7 CRC-enriched bacteria	NA	NA	0.800	Metagenomics	SVM	Feces	526	Meta	Dai et al., 2018
55 species	NA	NA	0.830	Metagenomics	RF	Feces	181	Meta	Sze and Schloss, 2018

DA, Decision Abundance; AR, Abundance Rate; SLC, Simple linear combination; Dirichlet MM, Dirichlet multinomial mixtures; Clostridium hathewayi (Ch); Unclassified species (m7); Roseburia intestinalis (Ri); Bacteroides thetaiotaomicron (Bt).

of Fn and FIT showed superior sensitivity to FIT alone, leading to detection of lesions missed by FIT alone (Wong et al., 2017a). Similarly, Pa, Pm, Cs, and m3 displayed an obvious improvement in both sensitivity and AUC, with a slight decrease in specificity (Xie et al., 2017; Liang et al., 2019). This complementary role was also illustrated using biomarker panels. Upon combining 22 OTUs identified using the penalized linear model with FIT, sensitivity increased from 58 to 72% at the same specificity (Zeller et al., 2014). In another study, combination of *Bacteroides clarus* (Bc), *Fn*, Ch, and m7 showed an increase of 9 percentage points when integrated with FIT in a logistic regression model (Liang et al., 2017). In conclusion, clinical screening programs based on both microbial markers and FIT/FOBT are cost-effective and present a promising diagnostic tool.

Prospects and Challenges

High-throughput sequencing and other analyses over the past decade have facilitated significant advances and gradual elucidation of the role of microbes in CRC. Current research on the value of clinical transformation of microbial markers in CRC diagnosis highlights the continued challenges of using available data effectively for making a contribution to precision medicine. Inspiration from other fields may additionally facilitate novel breakthroughs (**Figure 5**).

Formation of CRC is a multifactorial process and potential complementary effects between molecular markers require further attention. More than 80% CRC results from chromosomal instabilities, including mutation of the adenomatous polyposis coli (APC) gene and K-ras oncogene. APC gene-deficient mice can spontaneously grow tumors in the intestine and patients carrying the KRAS mutation show chemotherapeutic resistance (Colnot et al., 2004; Kuipers et al., 2015). Fecal DNA samples have been used to detect colorectal neoplasia (Imperiale et al., 2004). Septin 9 gene methylation has been shown to be effective as a biomarker and approved by the FDA (Lofton-Day et al., 2008). Meanwhile, methylation of bone morphogenic protein 3 and N-Myc downstream-regulated gene 4 displayed high specificity as an early and frequent event in colorectal tumors (Melotte et al., 2009; Loh et al., 2010). In 2014, multitarget stool DNA testing of combined KRAS, BMP3, NDRG4, and FIT achieved significantly higher detection of cancers, which led to FDA approval of Cologuard (Imperiale et al., 2014). Therefore, integration of genomics with microbiome analysis presents a promising direction. A recent study discussed this issue, suggesting that associations between tumor genomics and the microbiome could be beneficial in diagnostics (Burns and Blekhman, 2018). Since about 11% CRC cases result from overweight and obesity, other researchers used clinical data, such as body mass index (BMI) representing overall body fat, which displayed excellent discriminatory ability. However, no statistical significance was observed in a number of other analyses (Bardou et al., 2013; Zackular et al., 2014). To extract data from plain text files, Natural Language Processing methods or software have been employed for effective use of clinical features (Yim et al., 2016). Overall, these findings offer possible solutions and important directions for future research.

Universality is another key challenge, since differing opinions exist with regard to universal microbial markers. On the one hand, cross-cohort studies and meta-analyses have provided practicable and effective strategies that could overcome heterogeneity and ethnic differences with unbiased bioinformatics and statistical analysis. For instance, an earlier metagenomics analysis involving five ethnically different cohorts identified not only known biomarkers such as Fn,

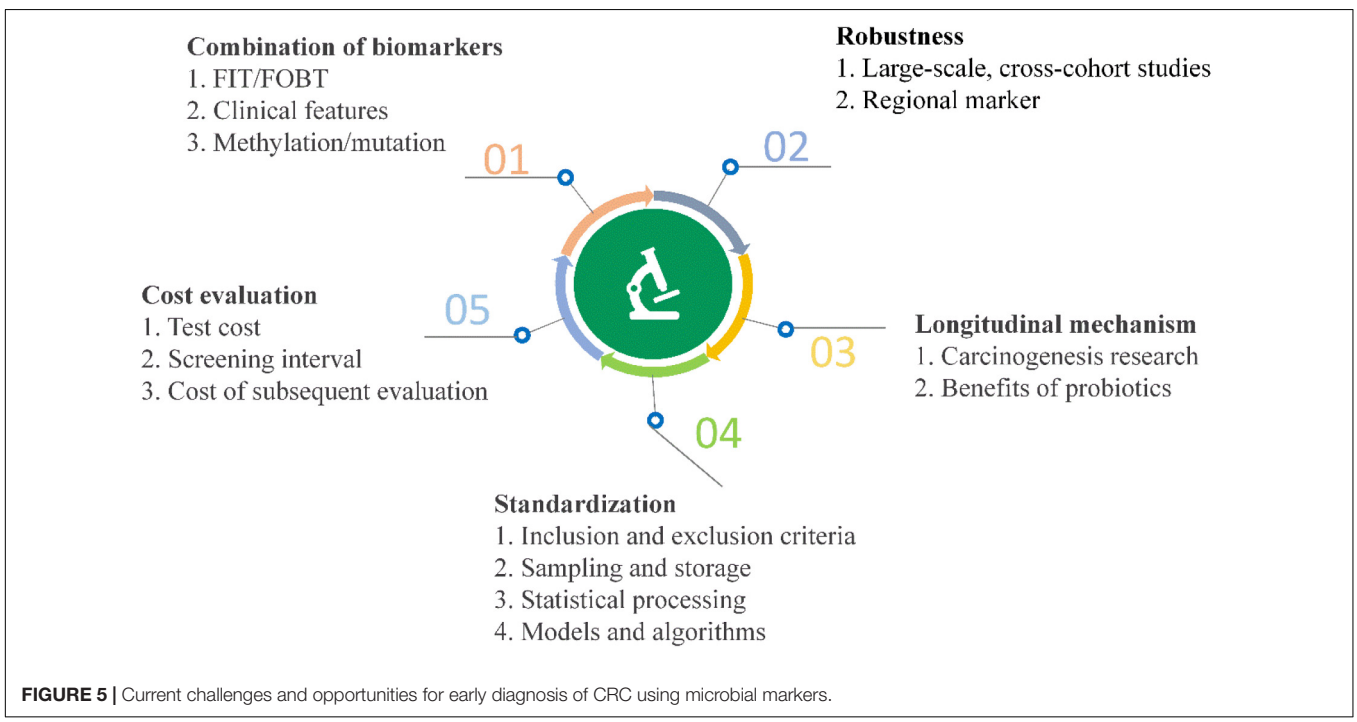


FIGURE 5 | Current challenges and opportunities for early diagnosis of CRC using microbial markers.

Ps, Pm, and *Solobacterium moorei*, but also a novel strain, *Peptostreptococcus anaerobius*, with subsequently confirmed roles in carcinogenesis using a $Apc^{Min/+}$ mouse model (Yu et al., 2017a; Long et al., 2019). Numerous meta-analyses also leveraged 16S rRNA or metagenomics data sets to reveal altered microbiome. Wirbel et al. (2019) identified a core set of 29 species while Dai et al. (2018) found 69 CRC-associated bacteria with metagenomic analysis. Similarly, two other teams identified 25 microbial OTUs and 12 common genera based on a random forest model using 16S rRNA sequencing datasets (Shah et al., 2018; Sze and Schloss, 2018). On the other hand, (Yang et al., 2020) proposed a strategy from a new angle, which inferred that regional biomarkers display high accuracy in specific populations. This theory was also supported by another study, which identified multiple *Fusobacterium* taxa (including *F. varium* and *F. ulcerans*) in Southern Chinese populations as disease biomarkers or targets that could be tailored according to discrepancies (Yeoh et al., 2020). Both alternative strategies provide well-powered assessments.

One of the significant challenges of clinical transformation is insufficient mechanistic analysis. While efficient computational frameworks and tools based on feature selection have been developed, machine learning requires further research (Tabib et al., 2020). Distinct from FIT/FOBT and fecal DNA tests, these semi-supervised or supervised learning methods are more like a "black box" with unclear mechanism. To date, hundreds of microorganisms have been shown to be linked with CRC, among which limited numbers have been further investigated. As a case in point, Fn was shown to be overabundant in tumor versus matched normal tissue and its potential role in CRC attracted widespread research attention (Castellarin et al., 2012; Kostic et al., 2012). Over the last few years, numerous studies have supported a role of Fn in promoting colorectal carcinogenesis through various functions such as inducing inflammatory cell infiltration, modulating E-cadherin/β-catenin signaling, activating immune cells, mediating interactions between bacteria, and binding to tumor-expressing Gal-GalNAc (Rubinstein et al., 2013, 2019; Abed et al., 2016; Yang et al., 2017). These advances have enhanced our knowledge of the potential relationships between Fn and chemoresistance, metastasis and poor prognosis (Mima et al., 2016; Yu et al., 2017b; Chen et al., 2020). Therefore, detection of Fn for early screening or exploitation of inhibitors targeting related pathways may be efficacious in clinical practice. In terms of methodological aspects, Bertrand Routy proposed a viable solution involving five steps: (1) microbial metagenomics should be standardized, (2) different "omics" analyses should be integrated, (3) the amount of cultivable microbial species should be increased, (4) non-invasive sampling methods should be combined with capsule endoscopy, and (5) Avatar mouse models should be standardized and investigated (Routy et al., 2018). Overall, longitudinal profiling of etiological and protection mechanisms of microorganisms achieves higher information richness and pave the way to take advantage of gut microbiome for diagnosis.

Development of standardized methods should also attenuate inconsistency of data. Inclusion and exclusion criteria have been gradually established, including diet, treatment, genetic background, disease history, antibiotic usage history and colonoscopy, aiming to avoid intestinal microbiota changes (O'Brien et al., 2013). During transportation and storage, a low temperature of −80°C and preservative buffer, such as RNAlater or EDTA, are effective to maintain DNA stability and integrity (Carozzi and Sani, 2013). In particular, compared to freezing for preservation, smaller technical variability was introduced without disrupting subject- and time-point specificity of the gut microbiome (Voigt et al., 2015). DNA extraction exerted the most significant effect on outcome of metagenomics analysis, highlighting the standardized DNA extraction method for human fecal samples (Costea et al., 2017). To address the complex challenges posed by large-scale studies, a protocol involving collection of microbiome samples at home and shipping to laboratories for molecular analysis was developed by Franzosa et al. (2014). Furthermore, for library preparation, PCR-free based methods were recommended to reduce PCR bias and improve assembly for accurate taxonomic assignment (Jones et al., 2015). Nevertheless, lack of standardization with regard to data access, metadata and analysis tools remain a barrier to acquisition of accurate and comparative results (Laudadio et al., 2018). Data integration and system-level modeling from multiple omics platforms is one of the most promising directions of microbiome research (Nayfach and Pollard, 2016). To improve the *status quo*, comprehensive platforms, such as MicrobiomeAnalyst and gcMeta, were recently constructed for downstream statistical analysis and functional interpretation (Dhariwal et al., 2017; Shi et al., 2019). Notably, the International Human Microbiome Standards (IHMS) project is committed to coordinate the development of standard operating procedures designed to optimize data quality and comparability in the human microbiome field. SYBR Green and probe-based qPCR are two common choices toward application, the former being more economical and the latter achieving greater accuracy for absolute quantification.

Cost-effectiveness is the ultimate challenge, including the costs of testing, screening intervals and subsequent evaluations resulting from the initial test (Dickinson et al., 2015). Due to high-cost resources, colonoscopy is not generally employed as a screening tool, except in a few countries like the United States, Germany and Austria. In low-income or middle-income countries with a low incidence of CRC, colonoscopy screening strategies may not be sufficiently cost-effective for implementation (Keum and Giovannucci, 2019). Taking FIT and Cologuard as examples, although incremental costs per additional advanced adenoma (AA) and CRC detected using colonoscopy versus FIT were £7,354 and £180,778, respectively, annual FIT reduced the colonoscopy incidence by 71% in intermediate-risk patients compared to three-yearly colonoscopy surveillance (Cross et al., 2019b). Cologuard shows superior performance for screening of AA, but carries a higher cost. In terms of the rate of screening compliance, stool DNA test is associated with higher patient acceptance owing to its simplicity. A preliminary calculation showed that combination of FIT and bacterial markers would avert up to 30% of total colonoscopies

as well as save an estimated 77 million € per 100,000 participants (Malagón et al., 2019). Meanwhile, usage of residual buffer from FIT cartridges is feasible for microbiota-based analysis and could greatly ameliorate the cost (Baxter et al., 2016a; Gudra et al., 2019).

Considering the collective findings, bacteriophages, viruses, archaea and fungi will be integrated into this database as biomarkers in the future. In addition, with advances in elucidation of mechanisms and omics analyses (such as transcriptomics, proteomics, and metabolomics), corresponding function descriptions should be more systematic. Systems biology and computational biology play crucial roles in mass data integration, and machine learning-based algorithms are under development for analysis of metadata to facilitate CRC diagnosis.

CONCLUSION

Development of colorectal cancer is a multifactorial process in which gut microbes play an important role. Determination of dysbiosis of microbial communities and differential patterns of abundance of microorganisms as biomarkers based on sequencing, algorithms and experimental data may aid in diagnosis and reduce morbidity and mortality. Except for a few pathogenic bacteria, the relationships between several microorganisms and colorectal cancer remain to be established, which are reflected by inconsistencies among different studies. Here, a database of CRC-related microbes was constructed using SQL and basic statistical analyses were conducted to outline biomarkers at different taxon levels. Diagnostic performance and mechanisms are discussed in detail. This type of knowledge integration is important for understanding and monitoring CRC. Moreover, this database can be used to perform inquiries and comparisons across different models and databases, contributing to further study of CRC-related microbes and promotion of cost-effective and non-invasive CRC screening strategies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

PC and ZZ contributed to the study design and drafted the manuscript. ZZ, SG, YaL, WM, YuL, SH, RZ, YM, KD and AS performed the statistical analysis and interpretation. All authors contributed to critical revision of the final manuscript and approved the final version of the manuscript.

FUNDING

This work was supported by Special Funding for Open and Shared Large-Scale Instruments and Equipments of Lanzhou University (Grant No. LZU-GXJJ-2019C012), the Project of Lanzhou City for Innovative and Entrepreneurial Talents (Grant No. 2017-RC-73), and Science and Technology Project of Lanzhou City (Grant No. 2018-4-59).

REFERENCES

Abed, J., Emgård, J. E., Zamir, G., Faroja, M., Almogy, G., Grenov, A., et al. (2016). Fap2 mediates *Fusobacterium nucleatum* colorectal adenocarcinoma enrichment by binding to Tumor-Expressed Gal-GalNAc. *Cell Host Microbe* 20, 215–225. doi: 10.1016/j.chom.2016.07.006

Arabameri, A., Asemani, D., and Teymourpour, P. (2020). Detection of colorectal carcinoma based on microbiota analysis using generalized regression neural networks and nonlinear feature selection. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 547–557. doi: 10.1109/tcbb.2018.2870124

Bardou, M., Barkun, A. N., and Martel, M. (2013). Obesity and colorectal cancer. *Gut* 62, 933–947. doi: 10.1136/gutjnl-2013-304701

Baxter, N. T., Koumpouras, C. C., Rogers, M. A. M., Ruffin, M. T., and Schloss, P. D. (2016a). DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. *Microbiome* 4:59. doi: 10.1186/s40168-016-0205-y

Baxter, N. T., Ruffin, M. T., Rogers, M. A. M., and Schloss, P. D. (2016b). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* 8:37. doi: 10.1186/s13073-016-0290-3

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Burns, M. B., and Blekhman, R. (2018). Integrating tumor genomics into studies of the microbiome in colorectal cancer. *Gut Microbes* 10, 547–552. doi: 10.1080/19490976.2018.1549421

Carozzi, F. M., and Sani, C. (2013). Fecal collection and stabilization methods for improved fecal DNA test for colorectal cancer in a screening setting. *J. Cancer Res.* 2013:818675. doi: 10.1155/2013/818675

Castellarin, M., Warren, R. L., Freeman, J. D., Dreolini, L., Krzywinski, M., Strauss, J., et al. (2012). *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res.* 22, 299–306. doi: 10.1101/gr.126516.111

Chen, Y., Chen, Y., Zhang, J., Cao, P., Su, W., Deng, Y., et al. (2020). *Fusobacterium nucleatum* promotes metastasis in colorectal cancer by activating autophagy signaling via the Upregulation of CARD3 expression. *Theranostics* 10, 323–339. doi: 10.7150/thno.38870

Chung, L., Thiele Orberg, E., Geis, A. L., Chan, J. L., Fu, K., DeStefano Shields, C. E., et al. (2018). *Bacteroides* fragilis toxin coordinates a pro-carcinogenic inflammatory cascade via targeting of colonic epithelial cells. *Cell Host Microbe* 23, 203–214.e5. doi: 10.1016/j.chom.2018.01.007

Colnot, S., Niwa-Kawakita, M., Hamard, G., Godard, C., Le Plenier, S., Houbron, C., et al. (2004). Colorectal cancers in a new mouse model of familial adenomatous polyposis: influence of genetic and environmental modifiers. *Lab. Invest.* 84, 1619–1630. doi: 10.1038/labinvest.3700180

Costea, P. I., Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., et al. (2017). Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* 35, 1069–1076. doi: 10.1038/nbt.3960

Cross, A. J., Wooldrage, K., Robbins, E. C., Kralj-Hans, I., MacRae, E., Piggott, C., et al. (2019a). Faecal immunochemical tests (FIT) versus colonoscopy for surveillance after screening and polypectomy: a diagnostic accuracy and cost-effectiveness study. *Gut* 68, 1642–1652. doi: 10.1136/gutjnl-2018-317297

Cross, A. J., Wooldrage, K., Robbins, E. C., Kralj-Hans, I., MacRae, E., Piggott, C., et al. (2019b). Faecal immunochemical tests (FIT) versus colonoscopy

for surveillance after screening and polypectomy: a diagnostic accuracy and cost-effectiveness study. *Gut* 68, 1642–1652. doi: 10.1136/gutjnl-2018-317297

Cuevas-Ramos, G., Petit, C. R., Marcq, I., Boury, M., Oswald, E., and Nougayrède, J. P. (2010). *Escherichia coli* induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 107, 11537–11542. doi: 10.1073/pnas.1001261107

Dai, Z., Coker, O. O., Nakatsu, G., Wu, W. K., Zhao, L., Chen, Z., et al. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* 6:70. doi: 10.1186/s40168-018-0451-2

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., and Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* 45, W180–W188. doi: 10.1093/nar/gkx295

Dickinson, B. T., Kisiel, J., Ahlquist, D. A., and Grady, W. M. (2015). Molecular markers for colorectal cancer screening. *Gut* 64, 1485–1494. doi: 10.1136/gutjnl-2014-308075

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* 8:1784. doi: 10.1038/s41467-017-01973-8

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591

Eklöf, V., Löfgren-Burström, A., Zingmark, C., Edin, S., Larsson, P., Karling, P., et al. (2017). Cancer-associated fecal microbial markers in colorectal cancer detection. *Int. J. Cancer* 141, 2528–2536. doi: 10.1002/ijc.31011

Eslami, M., Yousefi, B., Kokhaei, P., Hemati, M., Nejad, Z. R., Arabkari, V., et al. (2019). Importance of probiotics in the prevention and treatment of colorectal cancer. *J. Cell Physiol.* 234, 17127–17143. doi: 10.1002/jcp.28473

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. commun.* 6:6528. doi: 10.1038/ncomms7528

Flemer, B., Lynch, D. B., Brown, J. M., Jeffery, I. B., Ryan, F. J., Claesson, M. J., et al. (2017). Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* 66, 633–643. doi: 10.1136/gutjnl-2015-309595

Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., Das, A., Jeffery, I. B., et al. (2018). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* 67, 1454–1463. doi: 10.1136/gutjnl-2017-314814

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* 111, E2329–E2338. doi: 10.1073/pnas.1319284111

Gao, Z., Guo, B., Gao, R., Zhu, Q., and Qin, H. (2015). Microbiota disbiosis is associated with colorectal cancer. *Front. Microbiol.* 6:20. doi: 10.3389/fmicb.2015.00020

Gies, A., Cuk, K., Schrotz-King, P., and Brenner, H. (2018). Direct comparison of diagnostic performance of 9 quantitative fecal immunochemical tests for colorectal cancer screening. *Gastroenterology* 154, 93–104. doi: 10.1053/j.gastro.2017.09.018

Goodwin, A. C., Destefano Shields, C. E., Wu, S., Huso, D. L., Wu, X., Murray-Stewart, T. R., et al. (2011). Polyamine catabolism contributes to enterotoxigenic *Bacteroides* fragilis-induced colon tumorigenesis. *Proc. Natl. Acad. Sci. U. S. A.* 108, 15354–15359. doi: 10.1073/pnas.1010203108

Gudra, D., Shoaie, S., Fridmanis, D., Klovins, J., Wefer, H., Silamikelis, I., et al. (2019). A widely used sampling device in colorectal cancer screening programmes allows for large-scale microbiome studies. *Gut* 68, 1723–1725. doi: 10.1136/gutjnl-2018-316225

Guo, S., Li, L., Xu, B., Li, M., Zeng, Q., Xiao, H., et al. (2018). A Simple and novel fecal biomarker for colorectal cancer: ratio of *Fusobacterium Nucleatum* to probiotics populations. *Based on Their Antagonistic Effect. Clin. Chem.* 64, 1327–1337. doi: 10.1373/clinchem.2018.289728

Imperiale, T. F., Ransohoff, D. F., and Itzkowitz, S. H. (2014). Multitarget stool DNA testing for colorectal-cancer screening. *N. Engl. J. Med.* 371, 187–188. doi: 10.1056/NEJMc1405215

Imperiale, T. F., Ransohoff, D. F., Itzkowitz, S. H., Turnbull, B. A., and Ross, M. E. (2004). Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. *N. Engl. J. Med.* 351, 2704–2714. doi: 10.1056/nejmoa033403

Jones, M. B., Highlander, S. K., Anderson, E. L., Li, W., Dayrit, M., Klitgord, N., et al. (2015). Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc. Natl. Acad. Sci. U. S. A.* 112, 14024–14029. doi: 10.1073/pnas.1519288112

Keum, N., and Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat. Rev. Gastroenterol. Hepatol.* 16, 713–732. doi: 10.1038/s41575-019-0189-8

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* 22, 292–298. doi: 10.1101/gr.126573.111

Kosumi, K., Hamada, T., Koh, H., Borowsky, J., Bullman, S., Twombly, T. S., et al. (2018). The amount of bifidobacterium genus in colorectal carcinoma tissue in relation to tumor characteristics and clinical outcome. *Am. J. Pathol.* 188, 2839–2852. doi: 10.1016/j.ajpath.2018.08.015

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., et al. (2015). Colorectal cancer. *Nat. Rev. Dis. Primers* 1:15065. doi: 10.1038/nrdp.2015.65

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Kwong, T. N. Y., Wang, X., Nakatsu, G., Chow, T. C., Tipoe, T., Dai, R. Z. W., et al. (2018). Association between bacteremia from specific microbes and subsequent diagnosis of colorectal cancer. *Gastroenterology* 155, 383–390.e8. doi: 10.1053/j.gastro.2018.04.028

Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., and Carissimi, C. (2018). Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. *Omics* 22, 248–254. doi: 10.1089/omi.2018.0013

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023. doi: 10.1038/4441022a

Liang, J. Q., Chiu, J., and Chen, Y. (2017). Fecal bacteria act as novel biomarkers for non-invasive diagnosis of colorectal cancer. *Clin. Cancer Res.* 23, 2061–2070. doi: 10.1158/1078-0432.ccr-16-1599

Liang, J. Q., Li, T., Nakatsu, G., Chen, Y. X., Yau, T. O., Chu, E., et al. (2019). A novel faecal Lachnoclostridium marker for the non-invasive diagnosis of colorectal adenoma and cancer. *Gut* 69, 1248–1257. doi: 10.1136/gutjnl-2019-318532

Lofton-Day, C., Model, F., Devos, T., Tetzner, R., Distler, J., Schuster, M., et al. (2008). DNA methylation biomarkers for blood-based colorectal cancer screening. *Clin. Chem.* 54, 414–423. doi: 10.1373/clinchem.2007.095992

Loh, K., Chia, J., and Greco, S. (2010). Bone morphogenic protein 3 inactivation is an early and frequent event in colorectal cancer development. *Genes Chromosomes Cancer* 47, 449–460. doi: 10.1002/gcc.20552

Long, X., Wong, C. C., Tong, L., Chu, E. S. H., Ho Szeto, C., Go, M. Y. Y., et al. (2019). Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nat. Microbiol.* 4, 2319–2330. doi: 10.1038/s41564-019-0541-3

Malagón, M., Ramió-Pujol, S., Serrano, M., Serra-Pagès, M., Amoedo, J., Oliver, L., et al. (2019). Reduction of faecal immunochemical test false-positive results using a signature based on faecal bacterial markers. *Aliment. Pharmacol. Ther.* 49, 1410–1420. doi: 10.1111/apt.15251

Marchesi, J. R., and Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome* 3:31. doi: 10.1186/s40168-015-0094-5

Melotte, V., Lentjes, M. H. F. M., Van, d. B, and Sandra,, M. (2009). N-Myc Downstream-Regulated Gene 4 (NDRG4): a candidate tumor suppressor gene and potential biomarker for colorectal cancer. *J. Nat. Cancer Inst.* 101, 916–927. doi: 10.1093/jnci/djp131

Mima, K., Nishihara, R., Qian, Z. R., Cao, Y., Sukawa, Y., Nowak, J. A., et al. (2016). *Fusobacterium nucleatum* in colorectal carcinoma tissue and patient prognosis. *Gut* 65, 1973–1980. doi: 10.1136/gutjnl-2015-310101

Minot, S. S., and Willis, A. D. (2019). Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. *Microbiome* 7:110. doi: 10.1186/s40168-019-0722-6

Mira-Pascual, L., Cabrera-Rubio, R., Ocon, S., Costales, P., Parra, A., Suarez, A., et al. (2015). Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers. *J. Gastroenterol.* 50, 167–179. doi: 10.1007/s00535-014-0963-x

Nakatsu, G., Li, X., Zhou, H., Sheng, J., Wong, S. H., Wu, W. K., et al. (2015). Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat. Commun.* 6:8727. doi: 10.1038/ncomms9727

Nayfach, S., and Pollard, K. S. (2016). Toward accurate and quantitative comparative metagenomics. *Cell* 166, 1103–1116. doi: 10.1016/j.cell.2016.08.007

O'Brien, C. L., Allison, G. E., Grimpen, F., and Pavli, P. (2013). Impact of colonoscopy bowel preparation on intestinal microbiota. *PLoS One* 8:e62815. doi: 10.1371/journal.pone.0062815

O'Keefe, S. J. (2016). Diet, microorganisms and their metabolites, and colon cancer. *Nat. Rev. Gastroenterol. Hepatol.* 13, 691–706. doi: 10.1038/nrgastro.2016.165

Paramsothy, S., Nielsen, S., Kamm, M. A., Deshpande, N. P., Faith, J. J., Clemente, J. C., et al. (2019). Specific bacteria and metabolites associated with response to fecal microbiota transplantation in patients with ulcerative colitis. *Gastroenterology* 156, 1440–1454.e2. doi: 10.1053/j.gastro.2018.12.001

Pitot, H. C. (1993). The molecular biology of carcinogenesis. *Cancer* 72(3 Suppl.), 962–970.

Pleguezuelos-Manzano, C., Puschhof, J., Huber, A. R., van Hoeck, A., Wood, H. M., Nomburg, J., et al. (2020). Mutational signature in colorectal cancer caused by genotoxic pks(+) E. coli. *Nature* 580, 7269–7273. doi: 10.1038/s41586-020-2080-8

Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 579, 567–574. doi: 10.1038/s41586-020-2095-1

Poyet, M., Groussin, M., Gibbons, S. M., Avila-Pacheco, J., Jiang, X., Kearney, S. M., et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* 25, 1442–1452. doi: 10.1038/s41591-019-0559-3

Qin, J., Li, R., Raes, J., Arumugam, M., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Rastelli, M., Cani, P. D., and Knauf, C. (2019). The gut microbiome influences host endocrine functions. *Endocr. Rev.* 40, 1271–1284. doi: 10.1210/er.2018-00280

Rezasoltani, S., Sharafkhah, M., Asadzadeh Aghdaei, H., Nazemalhosseini Mojarad, E., Dabiri, H., Akhavan Sepahi, A., et al. (2018). Applying simple linear combination, multiple logistic and factor analysis methods for candidate fecal bacteria as novel biomarkers for early detection of adenomatous polyps and colon cancer. *J. Microbiol. Methods* 155, 82–88. doi: 10.1016/j.mimet.2018.11.007

Routy, B., Gopalakrishnan, V., Daillère, R., Zitvogel, L., Wargo, J. A., and Kroemer, G. (2018). The gut microbiota influences anticancer immunosurveillance and general health. *Nat. Rev. Clin. Oncol.* 15, 382–396. doi: 10.1038/s41571-018-0006-2

Rubinstein, M. R., Baik, J. E., Lagana, S. M., Han, R. P., Raab, W. J., Sahoo, D., et al. (2019). *Fusobacterium nucleatum* promotes colorectal cancer by inducing Wnt/β-catenin modulator Annexin A1. *EMBO Rep.* 20:e47638. doi: 10.15252/embr.201847638

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., and Han, Y. W. (2013). *Fusobacterium nucleatum* promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host Microbe* 14, 195–206. doi: 10.1016/j.chom.2013.07.012

Sacks, D., Baxter, B., Campbell, B. C. V., Carpenter, J. S., Cognard, C., Dippel, D., et al. (2018). Multisociety consensus quality improvement revised consensus statement for endovascular therapy of acute ischemic stroke. *Int. J. Stroke* 13, 612–632. doi: 10.1177/1747493018778713

Saffarian, A., Mulet, C., Regnault, B., Amiot, A., Tran-Van-Nhieu, J., Ravel, J., et al. (2019). Crypt- and mucosa-associated core microbiotas in humans and their alteration in colon cancer patients. *mBio* 10, e1315–e1319. doi: 10.1128/mBio.01315-19

Schloss, P. D., and Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbio.l* 77, 3219–3226. doi: 10.1128/aem.02810-10

Shah, M. S., Desantis, T. Z., Weinmaier, T., Mcmurdie, P. J., Cope, J. L., Altrichter, A., et al. (2018). Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut* 67, 882–891. doi: 10.1136/gutjnl-2016-313189

Shi, W., Qi, H., Sun, Q., Fan, G., Liu, S., Wang, J., et al. (2019). gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Res.* 47, D637–D648. doi: 10.1093/nar/gky1008

Sobhani, I., Tap, J., Roudot-Thoraval, F., Roperch, J. P., Letulle, S., Langella, P., et al. (2011). Microbial dysbiosis in colorectal cancer (CRC) patients. *PLoS One* 6:e16393. doi: 10.1371/journal.pone.0016393

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L. G., Gratadoux, J. J., et al. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U. S. A.* 105, 16731–16736. doi: 10.1073/pnas.0804812105

Sze, M. A., and Schloss, P. D. (2018). Leveraging Existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. *mBio* 9, e630–e618. doi: 10.1128/mBio.00630-18

Tabib, N. S. S., Madgwick, M., Sudhakar, P., Verstockt, B., Korcsmaros, T., and Vermeire, S. (2020). Big data in IBD: big progress for clinical practice. *Gut* 69, 1520–1532. doi: 10.1136/gutjnl-2019-320065

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019a). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019b). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7

Tilg, H., Adolph, T. E., Gerner, R. R., and Moschen, A. R. (2018). The intestinal microbiota in colorectal cancer. *Cancer Cell* 33, 954–964. doi: 10.1016/j.ccell.2018.03.004

Tjalsma, H., Boleij, A., Marchesi, J. R., and Dutilh, B. E. (2012). A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. *Nat. Rev. Microbiol.* 10, 575–582. doi: 10.1038/nrmicro2819

Viljoen, K. S., Dakshinamurthy, A., Goldberg, P., and Blackburn, J. M. (2015). Quantitative profiling of colorectal cancer-associated bacteria reveals associations between fusobacterium spp., enterotoxigenic *Bacteroides* fragilis (ETBF) and clinicopathological features of colorectal cancer. *PLoS One* 10:e0119462. doi: 10.1371/journal.pone.0119462

Villéger, R., Lopès, A., Veziant, J., Gagnière, J., Barnich, N., Billard, E., et al. (2018). Microbial markers in colorectal cancer detection and/or prognosis. *World J. Gastroenterol.* 24, 2327–2347. doi: 10.3748/wjg.v24.i22.2327

Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A. Y., Hercog, R., et al. (2016). Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS One* 11:e0155362. doi: 10.1371/journal.pone.0155362

Voigt, A. Y., Costea, P. I., Kultima, J. R., Li, S. S., Zeller, G., Sunagawa, S., et al. (2015). Temporal and technical variability of human gut metagenomes. *Genome Biol.* 16:73. doi: 10.1186/s13059-015-0639-8

Wang, L., Tang, L., Feng, Y., Zhao, S., Han, M., Zhang, C., et al. (2020). A purified membrane protein from Akkermansia muciniphila or the pasteurised bacterium blunts colitis associated tumourigenesis by modulation of CD8(+) T cells in mice. *Gut* 69, 1988–1997. doi: 10.1136/gutjnl-2019-320105

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25, 679–689. doi: 10.1038/s41591-019-0406-6

Wong, S. H., Kwong, T. N. Y., Chow, T. C., Luk, A. K. C., Dai, R. Z. W., Nakatsu, G., et al. (2017a). Quantitation of faecal Fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia. *Gut* 66, 1441–1448. doi: 10.1136/gutjnl-2016-312766

Wong, S. H., and Yu, J. (2019). Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat. Rev. Gastroenterol. Hepatol.* 16, 690–704. doi: 10.1038/s41575-019-0209-8

Wong, S. H., Zhao, L., Zhang, X., Nakatsu, G., Han, J., Xu, W., et al. (2017b). Gavage of fecal samples from patients with colorectal cancer promotes intestinal carcinogenesis in germ-free and conventional mice. *Gastroenterology* 153, 1621–1633.e6. doi: 10.1053/j.gastro.2017.08.022

Wong, S. H., Zhao, L., Zhang, X., Nakatsu, G., and Yu, J. (2017c). Gavage of fecal samples from patients with colorectal cancer promotes intestinal carcinogenesis in germ-free and conventional mice. *Gastroenterology* 153, 1621–1633. doi: 10.1053/j.gastro.2017.08.022

Wu, J., Li, Q., and Fu, X. (2019). *Fusobacterium nucleatum* contributes to the carcinogenesis of colorectal cancer by inducing inflammation and suppressing host immunity. *Transl. Oncol.* 12, 846–851. doi: 10.1016/j.tranon.2019.03.003

Wu, N., Yang, X., Zhang, R., Li, J., Xiao, X., Hu, Y., et al. (2013). Dysbiosis signature of fecal microbiota in colorectal cancer patients. *Microb. Ecol.* 66, 462–470. doi: 10.1007/s00248-013-0245-9

Xia, Y., Sun, J., and Chen, D.-G. (2018). *Statistical Analysis of Microbiome Data with R*. Singapore: Springer Singapore.

Xie, Y. H., Gao, Q. Y., and Cai, G. X. (2017). Fecal clostridium symbiosum for noninvasive detection of early and advanced colorectal cancer: test and validation studies. *Ebiomedicine* 25, 32–40. doi: 10.1016/j.ebiom.2017.10.005

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* 25, 968–976. doi: 10.1038/s41591-019-0458-7

Yang, J., Li, D., Yang, Z., Dai, W., Feng, X., Liu, Y., et al. (2020). Establishing high-accuracy biomarkers for colorectal cancer by comparing fecal microbiomes in patients with healthy families. *Gut Microbes* 13, 918–929. doi: 10.1080/19490976.2020.1712986

Yang, Y., Misra, B. B., Liang, L., Bi, D., Weng, W., Wu, W., et al. (2019). Integrated microbiome and metabolome analysis reveals a novel interplay between commensal bacteria and metabolites in colorectal cancer. *Theranostics* 9, 4101–4114. doi: 10.7150/thno.35186

Yang, Y., Weng, W., Peng, J., Hong, L., Yang, L., Toiyama, Y., et al. (2017). *Fusobacterium nucleatum* increases proliferation of colorectal cancer cells and tumor development in mice by activating toll-like receptor 4 signaling to nuclear factor-κB, and up-regulating expression of MicroRNA-21. *Gastroenterology* 152, 851–866.e24. doi: 10.1053/j.gastro.2016.11.018

Yeoh, Y. K., Chen, Z., Wong, M. C. S., Hui, M., Yu, J., Ng, S. C., et al. (2020). Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor. *Gut* 69, 1998–2007. doi: 10.1136/gutjnl-2019-319635

Yim, W. W., Yetisgen, M., Harris, W. P., and Kwan, S. W. (2016). Natural language processing in oncology: a review. *JAMA Oncol.* 2, 797–804. doi: 10.1001/jamaoncol.2016.0213

Youssef, O., Lahti, L., Kokkola, A., Karla, T., Tikkanen, M., Ehsan, H., et al. (2018). Stool microbiota composition differs in patients with stomach. *Colon, and Rectal Neoplasms. Dig. Dis. Sci.* 63, 2950–2958. doi: 10.1007/s10620-018-5190-5

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Liang, Q. Y., Qin, Y., et al. (2017a). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66, 70–78. doi: 10.1136/gutjnl-2015-309800

Yu, T., Guo, F., Yu, Y., Sun, T., Ma, D., Han, J., et al. (2017b). *Fusobacterium nucleatum* promotes chemoresistance to colorectal cancer by modulating autophagy. *Cell* 170, 548–563.e16. doi: 10.1016/j.cell.2017.07.008

Zackular, J. P., Rogers, M. A., Ruffin, M. T. T., and Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res. (Phila)* 7, 1112–1121. doi: 10.1158/1940-6207.Capr-14-0129

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 February 2021
doi: 10.3389/fmicb.2021.613109

Impact of Temporal pH Fluctuations on the Coexistence of Nasal Bacteria in an *in silico* Community

*Sandra Dedrick[1], M. Javad Akbari[1], Samantha K. Dyckman[1], Nannan Zhao[2], Yang-Yu Liu[2] and Babak Momeni[1]**

[1] Department of Biology, Boston College, Chestnut Hill, MA, United States, [2] Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, United States

To manipulate nasal microbiota for respiratory health, we need to better understand how this microbial community is assembled and maintained. Previous work has demonstrated that the pH in the nasal passage experiences temporal fluctuations. Yet, the impact of such pH fluctuations on nasal microbiota is not fully understood. Here, we examine how temporal fluctuations in pH might affect the coexistence of nasal bacteria in *in silico* communities. We take advantage of the cultivability of nasal bacteria to experimentally assess their responses to pH and the presence of other species. Based on experimentally observed responses, we formulate a mathematical model to numerically investigate the impact of temporal pH fluctuations on species coexistence. We assemble *in silico* nasal communities using up to 20 strains that resemble the isolates that we have experimentally characterized. We then subject these *in silico* communities to pH fluctuations and assess how the community composition and coexistence is impacted. Using this model, we then simulate pH fluctuations—varying in amplitude or frequency—to identify conditions that best support species coexistence. We find that the composition of nasal communities is generally robust against pH fluctuations within the expected range of amplitudes and frequencies. Our results also show that cooperative communities and communities with lower niche overlap have significantly lower composition deviations when exposed to temporal pH fluctuations. Overall, our data suggest that nasal microbiota could be robust against environmental fluctuations.

Keywords: microbial communities, variable environment, nasal microbiota, mathematical model, species interaction network, community ecology, coexistence

OPEN ACCESS

Edited by:
Dong-Yup Lee, Sungkyunkwan University, South Korea

Reviewed by:
Francesco Vitali, National Research Council (CNR), Italy
Seo-Young Park, Tufts University, United States

***Correspondence:**
Babak Momeni
momeni@bc.edu

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *01 October 2020*
Accepted: *18 January 2021*
Published: *10 February 2021*

Citation:
Dedrick S, Akbari MJ, Dyckman SK, Zhao N, Liu Y-Y and Momeni B (2021) Impact of Temporal pH Fluctuations on the Coexistence of Nasal Bacteria in an in silico Community. Front. Microbiol. 12:613109. doi: 10.3389/fmicb.2021.613109

INTRODUCTION

Resident microbes in the human nasal passage protect us from respiratory pathogens (Brugger et al., 2016; Man et al., 2017). Indeed, previous research shows the role of resident commensals in suppressing pathogens, such as *Staphylococcus aureus* (Uehara et al., 2000; Iwase et al., 2010; Bomar et al., 2016). Investigating how this microbial community is formed and maintained can therefore

provide powerful insights into microbiota-based therapies to prevent or treat infections. While such an investigation appears formidable in complex environments such as the gut microbiota, it is feasible for nasal microbiota. First, the nasal microbiota has relatively low diversity, with the majority of composition often attributed to 3∼8 species (Escapa et al., 2018). Second, the majority of these species are readily culturable aerobically *in vitro* under controlled environments (Kaspar et al., 2016; Escapa et al., 2018). Third, both the species and the nasal environment can be sampled relatively easily (Yan et al., 2013; Proctor and Relman, 2017). The combination of these factors makes the nasal microbiota a suitable choice for mechanistic studies of human microbiota and a gateway for more detailed studies of human-associated microbiota. Despite these advantages, community-level modeling of nasal microbiota has not been discussed adequately so far. A majority of existing work has focused on the biology of specific members of the nasal microbiota such as *Staphylococcus aureus* or *Streptococcus pneumoniae* because of their disease relevance (Regev-Yochay et al., 2004; Wertheim et al., 2004; Cespedes et al., 2005). Other reports have characterized and investigated the interactions among nasal microbes (Iwase et al., 2010; Bomar et al., 2016), but often with a focus on the interaction itself, and have only rarely involved the ecological consequences for the community (see Margolis et al., 2010; Yan et al., 2013; Krismer et al., 2017, for example).

Many factors, including interspecies interactions (Bomar et al., 2016; Brugger et al., 2016, 2020), the host immune system (Johannessen et al., 2012), and resource availability and access (Relman, 2012) can impact the nasal microbiota. However, all these factors take place in an environment that may fluctuate over time and vary in space. Previous investigations have revealed that the nasal passage is in fact very heterogeneous, both spatially and temporally (Proctor and Relman, 2017). In particular, pH fluctuations (in the range of 5.8–7.2, depending on the sampling site and time) were observed within the nasal passage (Washington et al., 2000; Hehar et al., 2001). Previous studies also demonstrate that temporal environmental fluctuations can transition the community to a different state (Abreu et al., 2020) or increase and support biodiversity (Eddison and Ollason, 1978; Grover, 1988; Abrams and Holt, 2002; Jiang and Morin, 2007; Kremer and Klausmeier, 2013). The explanation is often based on the temporal niche partitioning mechanism; i.e., environment variations creates additional niches and allow for more species to coexist (Chesson, 2000; Amarasekare, 2003). The purpose of our work is not to introduce a new theoretical framework for modeling microbial communities. Instead, we aim for a predictive mathematical model to study the impact of temporal pH fluctuations on the nasal microbiota composition. Other factors notwithstanding, we specifically ask whether, and when, incorporating temporal pH fluctuations is necessary to accurately predict compositional outcomes.

To answer the above question, we first characterize six nasal bacterial isolates as representative of members present in the nasal community. The rationale behind choosing these nasal bacteria was that (1) we can culture these strains reliably in the same cultivation medium and conditions in the lab; (2) covering different *Corynebacterium* and *Staphylococcus* species, these strains capture some of the natural diversity of microbiota (Escapa et al., 2018); and (3) Some interactions among these strains has already been identified (Brugger et al., 2020). For instance, *Corynebacterium* have been used to inhibit *S. aureus* colonization (Uehara et al., 2000; Kiryukhina et al., 2013) and *S. aureus* promotes the growth of *C. accolens* and gets inhibited by *C. pseudodiphtheriticum* (Yan et al., 2013). We then use *in vitro* communities constructed from nasal isolates to quantify the community response to temporal pH variations. Then, with parameters relevant to nasal microbiota, we use a phenomenological model to represent microbes and their interactions in an environment with a temporally fluctuating pH. Based on our empirical characterizations of nasal bacteria, we construct *in silico* examples of nasal microbiota and quantify their response to temporal pH fluctuations. Our simulation results suggest that temporal pH fluctuations do not have a major impact on the stable coexistence of nasal bacteria. The outline of our procedure to assess the impact of temporal pH fluctuations on nasal microbiota is shown in **Figure 1**.

MATERIALS AND METHODS

Nasal Bacterial Strains

Six strains used in this study were isolated from two healthy individuals and kindly shared with us by Dr. Katherine Lemon (**Table 1**). Interactions between some of these strains and other nasal bacteria has been studied recently (Brugger et al., 2020).

Cultivation Conditions and Medium *in vitro*

As growth medium, we have used a 10-fold dilution of the Todd-Hewitt broth with yeast extract (THY, at an initial pH of 7.2). We have diluted THY to create an environment closer to the nutrient richness of the nasal passage (Krismer et al., 2014). For collecting cell-free filtrates, cells were grown in 15 ml of media in sterile 50 ml Falcon tubes with loose caps exposed to the room atmosphere. For growth rate and carrying capacity characterizations, cells were grown in flat-bottom 96-well plates. All cultures were grown at 37°C with continuous shaking at 250 rpm.

Characterizing the pH Response of Nasal Isolates *in vitro*

To assess the response of nasal strains, we grew them in 10% THY after adjusting the pH within the biologically relevant range of 5.1 and 7.5 at 0.3 intervals (pH buffered with 10 g/l of MOPS). For each strain, we measured the growth rate at low population sizes (before nutrients become limiting or byproducts become inhibitory) and the final carrying capacity. These values were measured by growing replicates of each strain (typically 6 replicates) in 96-well microtiter plates incubated inside a Synergy Mx plate reader. Growth rate and carrying capacity were

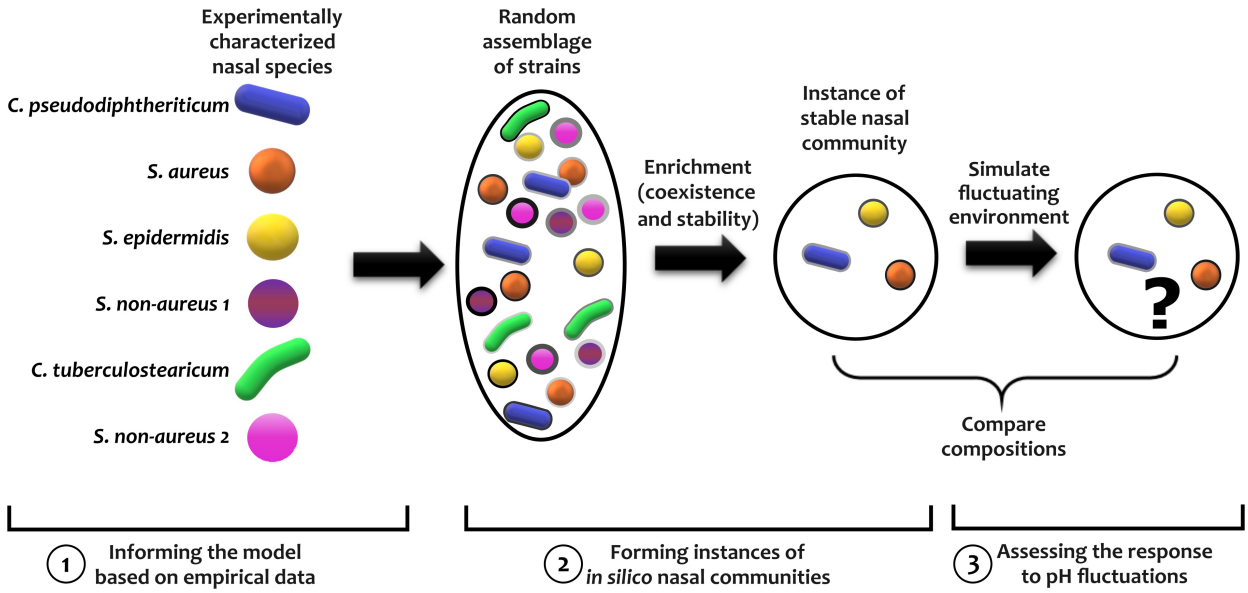


FIGURE 1 | The outline of the procedure for assessing the impact of temporal pH fluctuations on nasal microbiota is shown. To assemble *in silico* nasal communities, we characterized 6 bacterial nasal isolates. We then created *in silico* strains by randomly modulating the parameters of each characterized strain—schematically illustrated as different shades for each species. Using random assemblies of such strains, we simulated the enrichment process to find instances of stable nasal communities. We exposed these communities to a fluctuating pH and compared how the community composition was affected.

estimated by measuring the absorption in each well (OD_{600}) at 10 min intervals over 24 h at 37°C. Between absorption reads, the plate was kept shaking to ensure a well-mixed environment.

Mathematical Model

To model the growth of species, we assume that in the absence of interactions, the population growth follows the logistic equations:

$$\frac{dS_i}{dt} = r_i\,(p)\left[1 - \frac{S_i}{K_i\,(p)}\right]S_i - \delta S_i.$$

In which $r_i(p)$ and $K_i(p)$ are the pH-dependent growth rate and carrying capacity of species *i*. In our simulations, the growth rate and carrying capacity values at any given pH are found using a linear interpolation from experimentally measured values (pH 5.1–7.5 at 0.3 intervals). pH dependence is experimentally characterized for each strain in a monoculture, as described above, and δ is the dilution rate.

TABLE 1 | Nasal strains used in this study are listed along with their designation based on 16S rRNA gene similarity.

Strain name	Genus	Most likely species designation
KPL1821	*Corynebacterium*	*Corynebacterium tuberculostearicum*
KPL1828	*Staphylococcus*	*Staphylococcus aureus*
KPL1839	*Staphylococcus*	*Staphylococcus epidermidis*
KPL1850	*Staphylococcus*	*Staphylococcus non-aureus 1*
KPL1989	*Corynebacterium*	*Corynebacterium pseudodiphtheriticum*
KPL1867	*Staphylococcus*	*Staphylococcus non-aureus 2*

When multiple species are present, we assume that the presence of other species takes away resources from the environment; as a result, the growth of each species will be modulated as

$$\frac{dS_i}{dt} = r_i(p)\left[1 - \frac{S_i - \gamma_i}{K_i(p)}\right]S_i - \delta S_i.$$

where $\gamma_i = \sum_{j \neq i} c_{ij} S_j$ and $r_i(p)c_{ij}/K_i(p)$ represents the interaction strength exerted on species i by species j. Positive values of c_{ij} indicate growth stimulation (e.g., via facilitation by producing resources) whereas negative values of c_{ij} indicate growth inhibition (e.g., via competition).

Model Parameters

Unless otherwise specified, the following parameters are used in the model:

Some of these parameters, such as the range, frequency, and amplitude of pH values are chosen to keep the simulations close to what is expected in the nasal environment (Washington et al., 2000; Hehar et al., 2001). Some of the other parameters, such as the dilution rate or the initial and extinction population densities are not expected to be critical for the overall conclusions of this work. We have chosen these parameters to reflect realistic parameters that can be later tested experimentally. Finally, parameters such as the number of instances simulated (N_s) and the number of generations simulated (N_{gen}) are chosen to give us enough confidence for our claims, while keeping the practical considerations of simulation time and effort in mind.

Parameter	Description	Value
N_C	Maximum number of strains for *in silico* community assembly	20
N_S	Number of instances of assembly simulations run for each case	10,000
N_{gen}	Number of generations simulated to obtain stable resident communities; also the number of generations simulated to assess response to environmental fluctuations	100
pH_{rng}	Range of pH values (both in experiments and in simulations)	5.1–7.5
δ	Dilution rate	0.03–0.3 h^{-1}
N_{ext}	Extinction population density per species (OD)	10^{-6}
f_p	Inter-strain parameter variation within each species	20%
S_0	Average initial cell density per strain (OD)	10^{-4}
f_{pH}	Frequency of sinusoidal temporal pH fluctuations	1 h^{-1}
ΔpH	Amplitude of sinusoidal temporal pH fluctuations	0.5

Characterizing the Interspecies Interactions Using a Supernatant Assay

To characterize how species j affects the growth of other species i, we use a supernatant assay in which species j is grown to saturation, then all the cells are filtered out using a 0.22 μm filter (PVDF syringe filters from Thomas Scientific). The growth rate and carrying capacity of species i is then measured when growing in the supernatant taken from cultures of species j. This formulation allows us to use the experimentally measurable supernatant responses to formulate a dynamical model for mixed cultures of multiple species.

Assuming a Lotka-Volterra model, the presence of another species modulates the growth rate proportionally to the size of the interacting partner, i.e.,

$$\frac{dS_i}{dt} = r_i\left[1 - \frac{S_i - c_{ij}S_j}{K_i}\right](1 - \frac{S_i - c_{ij}S_j}{K_i})S_i.$$

Calculating the parameters obtained from the cell-free spent media (CFSM), the carrying capacity for species i is reached at population $S_{i,cc}$ level when growth rate becomes zero, thus

$$\left(1 - \frac{S_{i,cc} - c_{ij}K_j}{K_i}\right) = 0.$$

Therefore, the carrying capacity in the supernatant assay (K_{ij}) is

$$K_{ij} = S_{i,cc} = K_i + c_{ij}K_j.$$

And the interaction coefficient (c_{ij}, effect of species j on species i) can be calculated as

$$c_{ij} = \frac{K_{ij} - K_i}{K_j}.$$

In the particular that species i and j are similar (self-effect), we have $K_{ii} = 0$ and $c_{ii} = -1$. It should be noted that there are limitations in using a Lotka-Volterra model. Such models may not accurately represent microbial interactions (Momeni et al., 2017). Additionally, under certain conditions, the solutions will exhibit instability. Particularly for this latter case, we examine the situations under which "runaway" growth instability may happen. Consider two mutualistic populations:

$$\frac{dS_1}{dt} = r_1\left[1 - \frac{S_1 - c_{12}S_2}{K_1}\right]S_1 \text{ and } \frac{dS_2}{dt} = r_2\left[1 - \frac{S_2 - c_{21}S_1}{K_2}\right]S_2$$

Instability can happen when the carrying capacity terms fail to act as a negative feedback to bound the population. This can happen when

$$\left[1 - \frac{S_1 - c_{12}S_2}{K_1}\right] > 1 \text{ and } \left[1 - \frac{S_2 - c_{21}S_1}{K_2}\right] > 1$$

This happens when $S_1 < c_{12}S_2$ and $S_2 < c_{21}S_1$. Satisfying both of these inequalities requires that c_{12} and $c_{12}c_{21} > 1$, which means strong mutual facilitation. In our dataset, we do not have examples of mutual or cyclic facilitation and facilitation interaction terms are small, suggesting that instability is not expected in our simulations. Nevertheless, these conditions should be kept in mind for other datasets, especially those with strong facilitation between community members.

Calculating Community Composition Deviations

To compare community composition of a community that experienced pH fluctuation with that of the same community simulated at a fixed pH, we calculated the Bray-Curtis dissimilarity measure using the f_dis function (option "BC") in MATLAB. The necessary files to reproduce the analysis are included in the accompanied source codes[1].

Estimating the Impact of pH Fluctuations

We consider two extremes, when the fluctuations in pH are (1) much faster or (2) much slower than the population dynamics of community members. In both cases, for our formulation we define $c_{ii} = -1$ and use the simplified model of populations at different pH:

$$\frac{dS_i}{dt} = r_i(p)\left[1 + \frac{\widehat{\gamma_i}}{K_i(p)}\right]S_i - \delta S_i$$

and

$$\widehat{\gamma_i} = \sum c_{ij}S_j$$

Case 1. Fast pH fluctuations: To estimate how the community responds under a rapidly changing pH, we use the framework of the Wentzel–Kramers–Brillouin (WKB) approximation. For the general case of $\frac{dS}{dt} = r(t)S$, we split the population dynamics into two terms, the primary exponential term and an envelope

[1]https://github.com/bmomeni/temporal-fluctuations

function, E, for which $E(t) = e^{-r_0 t} S(t)$ and thus,

$$\frac{dE}{dt} = [r(t) - r_0] E$$

Using the WKB approximation E can be written using the expansion

$$E = \exp\left[\frac{1}{\varepsilon}\sum_{n=0}^{\infty} \varepsilon^n E_n(t)\right]$$

By inserting this expansion into the differential equation, we obtain

$$\left(\frac{1}{\varepsilon}\sum_{n=0}^{\infty} \varepsilon^n \frac{d}{dt} E_n(t)\right) \exp\left[\frac{1}{\varepsilon}\sum_{n=0}^{\infty} \varepsilon^n E_n(t)\right] =$$

$$(r(t) - r_0) \exp\left[\frac{1}{\varepsilon}\sum_{n=0}^{\infty} \varepsilon^n E_n(t)\right]$$

Thus

$$\frac{1}{\varepsilon}\sum_{n=0}^{\infty} \varepsilon^n \frac{d}{dt} E_n(t) = r(t) - r_0$$

Assuming sinusoidal changes in pH, $p(t) = p_0 + p_d \sin(2\pi f t)$, to the first order, the temporal changes in growth rate can be approximated as, $r(t) = r_0 + r_d \sin(2\pi f t)$. Therefore,

$$\frac{1}{\varepsilon}\sum_{n=0}^{\infty} \varepsilon^n \frac{d}{dt} E_n(t) \approx r_d \sin\left(2\pi f t\right).$$

In the limit that $\varepsilon \to 0$the first terms of expansion for E are obtained as

$$\frac{d}{dt} E_0(t) = 0, .$$

$$\frac{d}{dt} E_1(t) \approx r_d \sin(2\pi f t)$$

Since the continuous dilutions in our setup keeps the populations finite, E_0 does not affect the solution. The dominant term for E thus becomes E_1 and we have

$$E_1(t) \approx \frac{-r_d}{2\pi f} \cos\left(2\pi f t\right)$$

Aa a result,

$$E(t) \propto \exp\left[\frac{-r_d}{2\pi f} \cos\left(2\pi f t\right)\right]$$

Importantly, the magnitude of change in this equation drops inversely proportional to the frequency of pH fluctuations f. This means that the impact of pH fluctuations diminishes at high frequencies, consistent with our intuition that in this case the community dynamics are incapable of following the environmental fluctuations and only respond to the mean value.

Case 2. Slow pH fluctuations: In this case, we assume the quasi-static approximation, in which fluctuations are so slow that the community approaches its steady-state at each temporal value of pH. In this situation, assuming $\frac{dS_i}{dt} = 0$, we can rearrange the equation at steady-state as

$$\left[r_i(p) - \delta\right] S_i = -\frac{r_i(p)}{K_i(p)} S_i \sum c_{ij} S_j$$

Rearranging this, we get

$$\left[\delta - r_i(p)\right] \frac{K_i(p)}{r_i(p)} = \sum c_{ij} S_j.$$

This can be written in matrix form as

$$[C]\,\underline{S} = \underline{b},$$

where $[C]$ contains the interaction coefficients and $b_i = \left[\delta - r_i(p)\right] \frac{K_i(p)}{r_i(p)}$; underline in our notation designates a vector. Since the interaction matrix $[C]$ is pH-independent in our model, the change in composition within this quasi-static approximation can be expressed as

$$[C]\left(\underline{S(t)} - \underline{S_0}\right) = \underline{b(t)} - \underline{b_0},$$

or.

$$\underline{\Delta S(t)} = \left(\underline{S(t)} - \underline{S_0}\right) = [C]^{-1}\left(\underline{b(t)} - \underline{b_0}\right) = [C]^{-1}\,\underline{\Delta b(t)}$$

We make an additional simplifying assumption that K_i and r_i change similarly with pH. This leads to $\Delta b_i \approx (r_{i0} - r_i(t)) \frac{K_{i0}}{r_{i0}}$. This means that the magnitude of change in community composition is the same as the change in the growth rate of species, regardless of the frequency of fluctuations, under this regime.

Allowing pH-Dependent Interaction Coefficients

To examine how pH-dependent interaction coefficients may affect our results, we assumed that each interaction coefficient has a linear dependence on pH with a slope (per unit pH) randomly selected from a uniform distribution in the range of $[-m, m]$. In other words,

$$c_{ij}(p) = c_{ij}(p_0) + m(p - p_0),$$

where $p_0 = 7.2$ is the pH at which our characterization is performed. We examined how the community composition deviated from the reference with a fixed pH, as m (and thus the pH-dependency) increased.

RESULTS

In vitro Characterization of Nasal Bacteria

We experimentally characterized how six representative nasal bacterial strains respond to different pH values in their environment. These bacterial strains were chosen from a set of isolates (see section "Materials and Methods") based on three major considerations: (1) they reliably grow

in our cultivation media under an aerobic environment; (2) they include commonly observed *Staphylococcus* and *Corynebacterium* species; and (3) they span the phylogenetic landscape of both closely and distantly related bacteria found in the nasal environment (Escapa et al., 2018). We assumed that each of these characterized strains is a representative strain of the corresponding species.

We first characterized the pH response of each strain by growing them under different environmental pH values. Different strains exhibited different degrees of pH dependency in their growth rates and carrying capacities (**Figure 2** and **Supplementary Figures S1, S2**). Among these strains, *S. epidermidis*, *S. non-aureus 1*, and *S. non-aureus 2* show fairly similar growth properties. We chose to treat these as separate species in our investigation, because—as shown later—they had considerably different interactions with other species (**Figure 3**).

We then examined how different species interact with one another. For this, we grew each species to its stationary phase in a monoculture, filtered out the cells, and measured how other strains grew in the resulting cell-free filtrates (see section "Materials and Methods"; similar to De Vos et al., 2017; Brugger et al., 2020). From these measurements, we estimated the interspecies interaction coefficients based on the generalized Lotka-Volterra model (**Figure 3**; see section "Materials and Methods"). In this formulation, baseline competition with complete niche overlap will result in an interaction coefficient of −1. Of note, from our experimental data we cannot distinguish the relative contribution of competitive niche overlap and interspecies facilitation. Nevertheless, for simplicity we only use "facilitation" for extreme cases in which facilitation outweighs competition and the interaction coefficient turns positive. We interpret different gradations of negative interaction coefficients from −1 to 0 as different degrees of niche overlap (with −1 indicating complete niche overlap), and cases with interaction coefficients less than −1 indicate inhibition beyond competition for resources. Among the 30 pairwise interaction coefficients, there were 3 positive values (bright blue, marked by "+" in **Figure 3**). For simplicity, throughout this manuscript, we assume that these interaction coefficients are not pH-dependent.

In silico Assembly of Nasal Bacterial Communities

To capture some of the diversity of nasal microbiota, we propose that other *in silico* strains of each species can be constructed by randomly modulating the measured properties of that species (i.e., growth rate, carrying capacity, and interaction coefficients). We chose the degree of strain-level modulation to be up to 20%, as a balance between intraspecies and interspecies diversity (**Supplementary Figure S3**).

To assess the response of nasal microbiota to temporal fluctuations in the environment, we first construct an ensemble of *in silico* communities that represent a subset of possible nasal communities. This is chosen as an alternative to performing an *in vivo* study, because performing these experiments with human subjects is not feasible and there is no reliable animal model for human nasal bacteria. Our approach is, in essence, similar to several other previous work that have used simple models to describe the dynamics of human-associated microbiota (Stein et al., 2013; Fisher and Mehta, 2014; Song et al., 2014; De Vos et al., 2017; Venturelli et al., 2018). Compared to *in vitro* studies, these *in silico* communities give us full control over confounding factors and allows us to examine the mechanisms contributing to sensitivity to pH fluctuations (Momeni et al., 2011). To construct *in silico* communities, we mimicked enrichment experiments (Goldford et al., 2018; Niehaus et al., 2019) by simulating the dynamics of an initial assemblage of 20 strains (sampled from the space of *in silico* strains) until the community reached stable coexistence. These *in silico* communities were largely robust against experimental noise in characterization (**Supplementary Figure S4**). The interspecies interactions in our model appear to be instrumental in the assembly of these *in silico* communities, as evidenced by changes when we assigned the interaction coefficients at given levels (**Supplementary Figure S5A**) or modulated the measured interactions (**Supplementary Figure S5B**). To assess how pH fluctuations in the environment influence nasal communities, we take several instances of *in silico* nasal communities, expose them to a fluctuating pH, and quantify how the community composition is affected. The entire process is outlined in **Figure 1**.

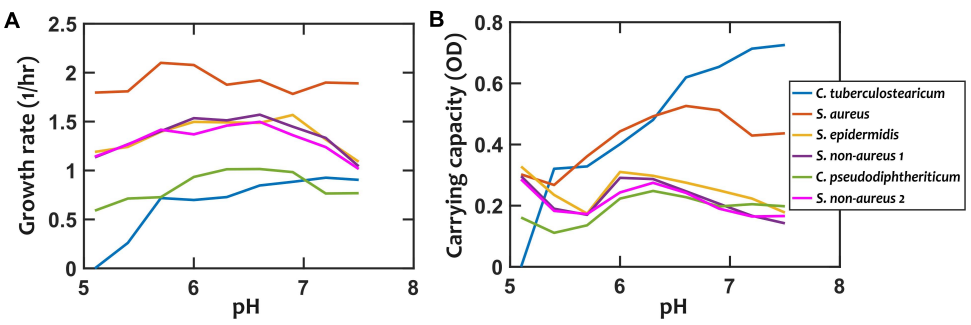


FIGURE 2 | Growth properties of nasal bacterial isolates are pH-dependent. Growth is characterized using the growth rate in the early exponential phase **(A)**, and the carrying capacity based on optical density (OD, absorption measured at 600 nm) as a proxy **(B)**. Each data point is the average of at least 6 replicates from two independent experiments. Error-bars are not shown to avoid overcrowding the plot but the values are available in the raw data. In all cases, growth is experimentally tested in a 10-fold diluted Todd-Hewitt broth with yeast extract (10% THY).

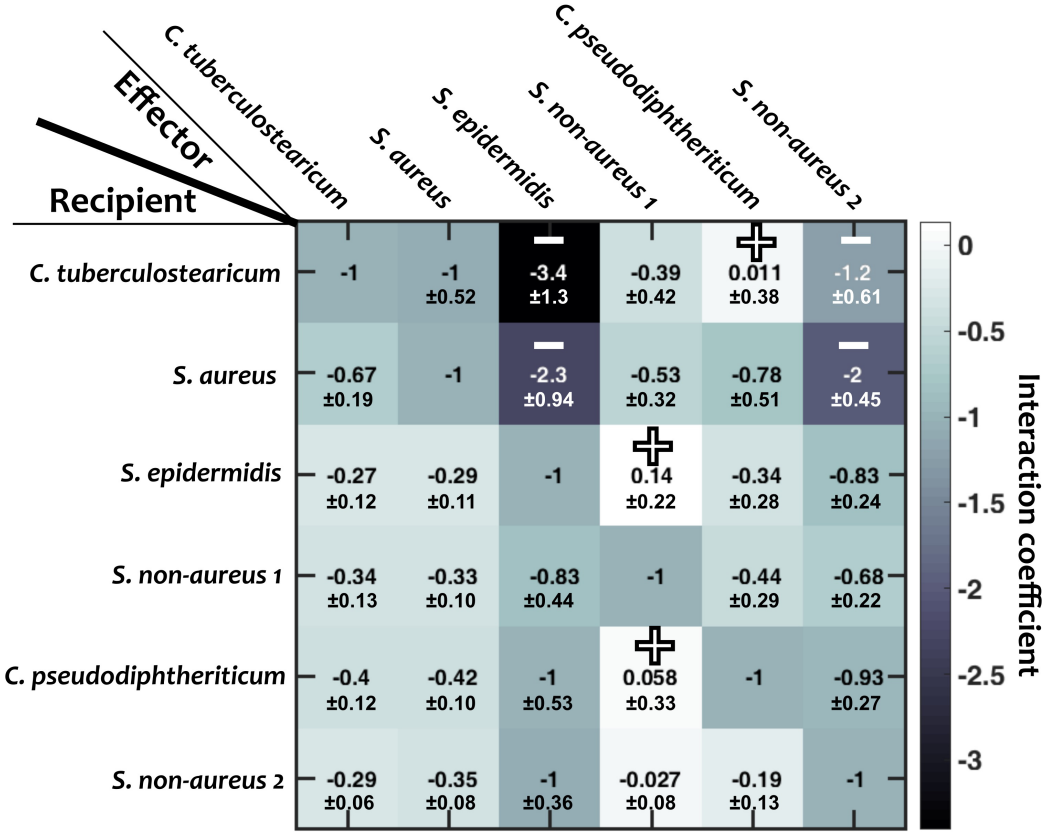


FIGURE 3 | Interaction coefficients among pairs of nasal bacteria. Values represent interaction coefficients in a Lotka-Volterra model. In each case, the growth of a recipient strain is measured when the strain is exposed to cell-free filtrate derived from the effector strain. Positive mean coefficients (indicating facilitation) and negative mean coefficients below –1.2 (indicating strong inhibition) across different replicates are marked by "+" and "–," respectively. Standard deviations (shown for each value) are calculated based on empirical standard deviations of measured carrying capacities in monocultures and supernatant experiments. Diagonal elements are set to –1, indicating complete niche overlap.

In silico Nasal Communities Are Diverse and Favor Facilitation

We first examined the properties of assembled *in silico* communities at various pH values with no temporal fluctuations. We found that the prevalence of different species was distinct and pH-dependent (**Supplementary Figure S6**). This prevalence is a result of nasal species' pH-dependent growth properties as well as their interspecies interactions.

We also found that during the process of assembling *in silico* communities, the prevalence of interspecies facilitation interactions increased. Comparing the prevalence of facilitation in initial assemblages of strains vs. the final stable communities, we found that among the communities that had at least one facilitation interaction at the start of the *in silico* enrichment (89% of communities), facilitation was enriched in ∼66% of the final community assemblies (**Supplementary Figure S7**).

Temporal pH Fluctuations Only Minimally Impact Nasal Microbiota Composition

Next, we asked how the temporal variation in the environment might influence the community composition. To answer this question, we used instances of *in silico* communities to evaluate the impact of temporal pH fluctuations. We assumed a continuous growth situation in which all community members experience a constant dilution rate. This dilution mimics the turnover in microbiota, for example, when the mucosal layer gets washed away. To avoid situations in which the *in silico* community itself was not stable, we changed the dilution rate by ± 50% and only kept the communities for which the modified dilutions only caused a small deviation in community composition (see section "Materials and Methods"). Indeed, we found that communities with compositions more sensitive to dilution rates are also more sensitive to pH fluctuations (**Supplementary Figure S8**). In all cases, composition deviations were calculated using the Bray-Curtis dissimilarity measure (see section "Materials and Methods").

To evaluate the impact of pH fluctuations, we simulated a controlled sinusoidal pH variation over time, with two parameters: the amplitude and frequency of temporal variations. Thus, $p(t) = p_0 + \Delta\text{pH}\sin(2\pi f_{\text{pH}}t)$. Keeping the frequency of fluctuations fixed ($f_{\text{pH}} = 0.2$/h), we observed that the deviation in population composition increased with an increasing pH fluctuation amplitude (ΔpH). However, the resulting

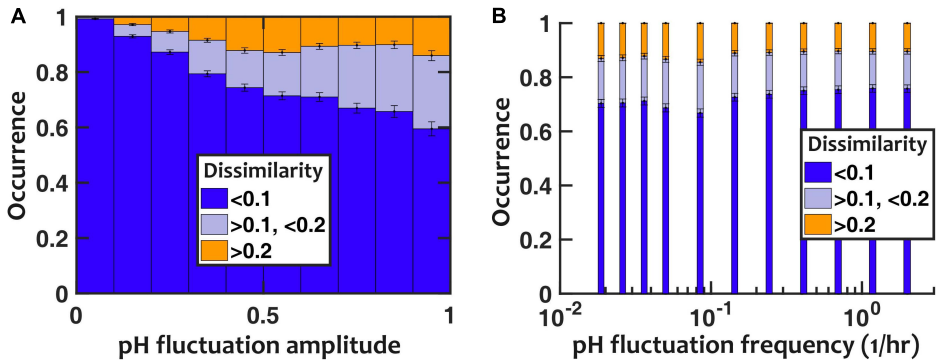


FIGURE 4 | Nasal microbiota composition is robust against pH fluctuations. **(A)** For a fixed fluctuation frequency (f_{pH} = 0.2/h), larger fluctuation amplitudes increase how the community composition deviates from the no-fluctuation steady state (as quantified with composition dissimilarity). **(B)** The impact of temporal fluctuations is maximum at intermediate frequencies. Here, the pH fluctuation amplitude is fixed (ΔpH = 0.5). Number of *in silico* communities examined for each condition: n = 10,000.

dissimilarity in population composition was mostly minor, with > 85% of cases showing less than 0.2 dissimilarity even when the amplitude of pH fluctuation was set to 1 (**Figure 4A**). We then examined the impact of the frequency of pH variations, while we kept the amplitude of pH fluctuations fixed (ΔpH = 0.5). At intermediate frequencies, the pH fluctuations caused the largest dissimilarity in community composition compared to stable communities with fixed pH (**Figure 4B**).

We repeated the assessment of pH fluctuations by assuming a pH that randomly fluctuated between two discrete pH values to ensure that our results were not limited to sinusoidal fluctuations. The results were overall consistent with sinusoidal pH fluctuations (**Supplementary Figure S9**): (1) larger pH fluctuation amplitudes increased the deviation in population composition, but overall the majority of communities only experienced modest deviations; and (2) pH fluctuations at intermediate frequencies had the largest impact on community composition.

Interspecies Facilitations Dampen the Impact of Temporal Fluctuations

To explain the low sensitivity of community composition to pH fluctuations, we hypothesized that interspecies facilitation stabilizes the composition by creating interdependencies within the community. From the data in **Figure 4**, we picked and compared communities with low ("competitive," 0% facilitation) and high ("cooperative," 50% facilitation) prevalence of facilitation. The 0% facilitation corresponds to situations where none of the strains facilitate any of the other members of the *in silico* community. In contrast, 50% facilitation happens when half of the pairwise interactions among the *in silico* community are facilitative. Since in our dataset (**Figure 3**) there were no instances of mutual facilitation, 50% facilitation is the maximum fraction that cooperative communities can reach. The results show that cooperative communities have a consistently and significantly lower composition deviation when exposed to temporal pH fluctuations (**Figure 5**).

To further explore the impact of facilitation, we asked how interspecies niche overlap (the magnitude of negative interspecies interactions) and prevalence of facilitation (the fraction of interspecies interactions that are positive) contribute to sensitivity to pH fluctuations. In our results, we found that larger interspecies niche overlap leads to more sensitivity to pH fluctuations (**Figure 6A**). This trend holds except when interspecies niche overlap approaches 1; at such high overlaps the community loses diversity (**Figure 6B**), becoming less sensitive to pH fluctuations. When we directly changed the prevalence of facilitation, we observed that with higher prevalence of facilitation the communities became more diverse and less sensitive to pH fluctuations (**Figures 6C,D**).

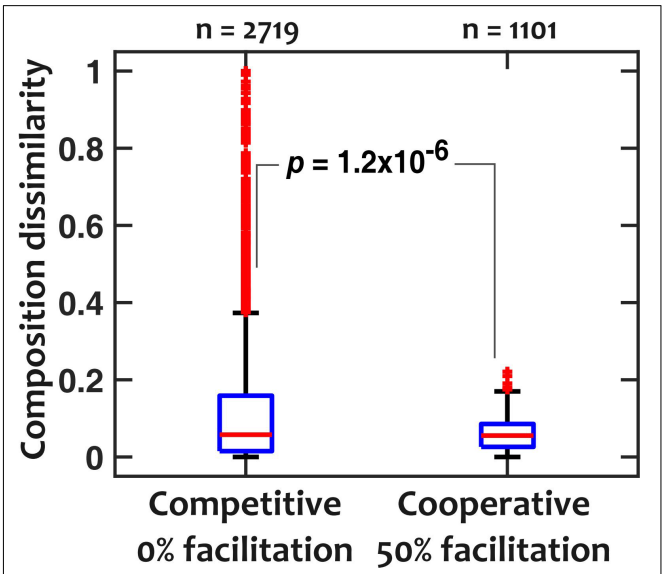


FIGURE 5 | Cooperative communities are more robust against pH fluctuations compared to competitive communities. For competitive (those with 0% facilitation among members) and cooperative (those with 50% facilitation interactions among members) communities, dissimilarity medians were 0.057 and 0.055 and dissimilarity means were 0.12 and 0.058, respectively ($p = 1.2 \times 10^{-6}$ with a Mann-Whitney U-test). pH fluctuates sinusoidally with a frequency of f_{pH} = 0.2/h and an amplitude of ΔpH = 0.5.

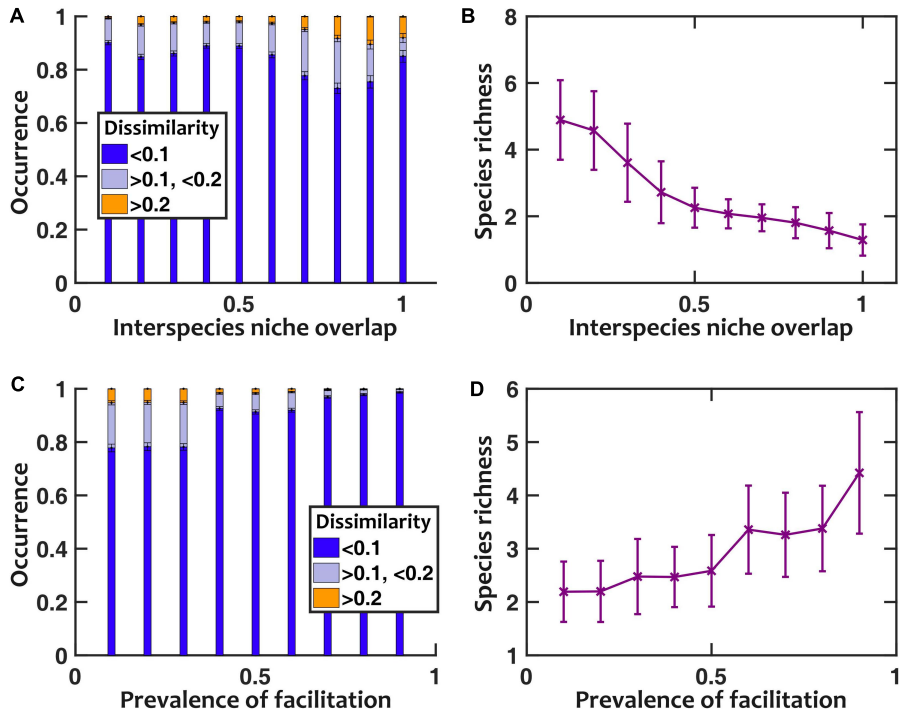


FIGURE 6 | Lower niche overlap and more prevalent facilitation decrease the sensitivity to pH fluctuations. **(A)** As we artificially increased the strength of niche overlap (by setting all off-diagonal coefficients in the interaction matrix to be a fixed negative number and making that number more negative), the sensitivity to pH fluctuations increased. This trend is disrupted when niche overlap approaches 1, because a majority of communities under such conditions lose interspecies diversity **(B)**. As we increased the prevalence of interspecies facilitation by randomly setting a given fraction of interaction coefficients to be positive, the communities became less sensitive to pH fluctuations **(C)** and more diverse in species richness **(D)**. Here pH fluctuations are sinusoidal, with f_{pH} = 0.2/h and ΔpH = 0.5.

DISCUSSION

Using empirically measured species properties, we assembled stable *in silico* communities that show coexistence of nasal bacteria. When these communities were exposed to a fluctuating pH environment, we observed that the composition of stable communities was only modestly affected. Larger pH fluctuations increased the deviation, as expected; however, even at a pH fluctuation of 1 which exceeds the observed temporal variation in the nasal passage, the composition of the majority of communities remained minimally affected. We also found that intermediate frequencies of temporal pH fluctuations caused the largest deviations in community compositions. Finally, in our results, communities with more facilitation interactions were more robust against pH fluctuations.

In choosing an appropriate model, one must carefully consider the processes of interest and the required level of abstraction to capture those processes (Momeni et al., 2011; Silverman et al., 2018b). Models originally designed for single species populations have been adapted to characterize microbial communities. Community ecology modeling frameworks designed to understand interactions at the macro scale—in both space and time—have also been applied to microbial populations to study the dynamics of succession and restoration, along with the impact of environmental disturbances (Byrd and Segre, 2016; Gilbert and Lynch, 2019). For example, flux balance analyses, a mainstay in microbial metabolic models, can be modified to describe species interactions within a complex microbial community over time (Larsen et al., 2012; Gerber, 2014; Bucci et al., 2016; Fukuyama et al., 2017; Äijö et al., 2018; Silverman et al., 2018a; Shenhav et al., 2019). To create a predictive model for nasal microbiota, we have extended the generalized Lotka-Volterra (gLV) equations to study the impact of pH fluctuations on community composition. Generalized Lotka-Volterra equations have been previously used to investigate species interactions in the human gut (Fisher and Mehta, 2014; Song et al., 2014) and in a cheese-associated microbial community (Äijö et al., 2018; Song et al., 2014). It has also been similarly extended to describe the impact of environmental fluctuations (antibiotics) on gut microbiota (Stein et al., 2013; Song et al., 2014). For our data, the Lotka-Volterra-type model has proven to—at least to the first-order—capture species growth, interactions, and pH-dependence.

One important aspect of temporal fluctuations is their time scale. Even though in nature the fluctuations are not completely regular, our investigation with sinusoidal temporal fluctuations reveal the time scale at which the influence on community composition is the strongest. Our analysis reveals that the fluctuations are more impactful at intermediate frequencies between two extremes (see section "Materials and Methods"). At very low frequencies of pH fluctuations, the community dynamics are faster than pH changes; thus we can assume

the quasi-static approximation applies. In this regime, the community reaches its stable state locally (in time), and the composition follows the value of pH at any given time, regardless of the frequency of the pH fluctuations. In the other extreme, at very high frequencies of pH fluctuations, the population dynamics cannot follow rapid changes in pH and essentially the species "see" the average pH. An analysis based on the Wentzel–Kramers–Brillouin (WKB) approximation suggests that in this regime, the magnitude of change in composition (compared to the composition at the average pH) is inversely proportional to the pH fluctuation frequency. Between these two extremes is the zone that exhibits the most change in community composition with pH fluctuations (**Figure 4B** and **Supplementary Figure S9B**). However, for parameters relevant to the nasal strains we are analyzing, even in this zone the changes in community composition are not drastic.

Our focus in this manuscript is on how composition of stable communities changes when environmental pH fluctuates. Another relevant question is how fluctuations in pH affect the process of community assembly. For this, we repeated the community assembly simulations (**Supplementary Figures S6,S7**), but under an environment in which the pH temporally fluctuated. Contrary to our expectation, the richness of resulting communities did not monotonically increase with an increase in the amplitude of pH fluctuations, regardless of fluctuation frequency (**Supplementary Figure S10**). Instead, we found that richness only changed in a small fraction of *in silico* nasal communities. Furthermore, in cases with increased richness, *S. non-aureus* 1 (most facilitative species in our panel) was most frequently added to the community, whereas in cases with decreased richness, *S. epidermidis* (most inhibitory species in our panel) was most frequently dropped from the community (**Supplementary Figure S11**). This observation underscores the relative importance of interaction (compared to niche partitioning) in richness outcomes in our model of nasal communities. Our finding is also consistent with predictions about augmentation and colonization resistance using a mediator-explicit model of interactions (Kurkjian et al., 2020).

There are some limitations and simplifications in our study. First, in our investigation we have assumed that fluctuations in pH are imposed externally (e.g., by the host or the environment). It is also possible that species within the nasal community contribute to the environmental pH. Although outside the scope of this work, we speculate that if species within the community drive the pH to specific values (Ratzke and Gore, 2018; Ratzke et al., 2018), the impact of external temporal fluctuations of pH on community composition will be even more diminished. Second, in our model, we assumed that interactions among species remained unchanged at different environmental pH values. We examined *in silico* how pH-dependent interaction coefficients might affect our results. For this, we assumed that interaction coefficients changed linearly with pH in each case (see section "Materials and Methods") and asked how strong the dependency had to be to considerably change the community composition under a fluctuating pH. We observed a significant impact only when the interaction coefficients were strongly pH dependent (i.e., to the level that the sign of interactions would change within the range of pH fluctuations) (**Supplementary Figure S12**).

Our work suggests that a shift in pH can change the community composition and coexistence (**Supplementary Figure S6**). This is consistent with previous observations from profiling different locations along the nasal passage (Yan et al., 2013). However, our prediction is that temporal pH fluctuations often do not cause a major shift in community structure. As a future step, we plan to verify this prediction experimentally by testing how pH fluctuations affect *in vitro* nasal communities. If confirmed, our prediction is that the spatial position of sampling and the heterogeneity of the environment will have a stronger effect on community composition compared to the temporal resolution of sampling. The practical implication is that microbiome profiling of nasal microbes may not require a high temporal resolution. We proposed that high throughput sampling of the nasal microbiome along with the corresponding pH would be an insightful future step to test our predictions.

Finally, one of the main messages of our work is that nasal microbiota is insensitive to temporal fluctuations in pH. It is tantalizing to speculate, when examining other microbial communities, under what conditions this statement is valid. Recent work by Shibasaki et al. (2020) shows that under a fluctuating environment species properties play an important role in community diversity. Our results corroborate their finding. Insensitivity of the members to the environmental fluctuations—as trivial as it may sound—is a defining factor for how sensitive the community is. In the nasal microbiota, species that we are examining are adapted to the nasal environment and the range of pH fluctuations experienced in this environment is not large. As a result, the community is not majorly affected by pH fluctuations. On top of this, we also observe that interactions—in particular, facilitation and competition—can act as stabilizing or de-stabilizing factors for how the community responds to external variations. In other words, facilitation between community members acts as a composition stabilizing factor between populations, which lowers the impact of external fluctuations (**Figure 6A**). In contrast, inhibition between community members typically exaggerates the changes introduced by external fluctuations (**Figure 6C**).

DATA AVAILABILITY STATEMENT

Codes related to this manuscript can be found at: https://github.com/bmomeni/temporal-fluctuations. The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SD, NZ, Y-YL, and BM conceived the research. SD and BM designed the simulations and experiments and wrote the manuscript. SD ran the experiments. SD, MA, SKD, NZ, and BM ran the simulations. SD, MA, SKD, NZ, Y-YL, and BM edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

SD was supported by the NIH T32 training grant. Work in the Momeni Lab was supported by a start-up fund from Boston College and by an Award for Excellence in Biomedical Research from the Smith Family Foundation. NZ was supported by the China Scholarship Council. Y-YL acknowledges grants R01AI141529, R01HD093761, UH3OD023268, U19AI095219, and U01HL089856 from National Institutes of Health.

ACKNOWLEDGMENTS

We would like to thank Dr. Katherine Lemon for kindly sharing the nasal bacterial strains with us. This manuscript has been released as a pre-print at (Dedrick et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.613109/full#supplementary-material

REFERENCES

Abrams, P. A., and Holt, R. D. (2002). The impact of consumer-resource cycles on the coexistence of competing consumers. *Theor. Popul. Biol.* 62, 281–295. doi: 10.1006/tpbi.2002.1614

Abreu, C. I., Andersen Woltz, V. L., Friedman, J., and Gore, J. (2020). Microbial communities display alternative stable states in a fluctuating environment. *PLoS Comput. Biol.* 16:e1007934. doi: 10.1371/journal.pcbi.1007934

Äijö, T., Müller, C. L., and Bonneau, R. (2018). Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics* 34, 372–380. doi: 10.1093/bioinformatics/btx549

Amarasekare, P. (2003). Competitive coexistence in spatially structured environments: a synthesis. *Ecol. Lett.* 6, 1109–1122. doi: 10.1046/j.1461-0248.2003.00530.x

Bomar, L., Brugger, S. D., Yost, B. H., Davies, S. S., and Lemon, K. P. (2016). *Corynebacterium accolens* releases antipneumococcal free fatty acids from human nostril and skin surface Triacylglycerols. *mBio* 7:e01725-15. doi: 10.1128/mBio.01725-1715

Brugger, S. D., Bomar, L., and Lemon, K. P. (2016). Commensal-pathogen interactions along the human nasal passages. *PLoS Pathog.* 12:e1005633. doi: 10.1371/journal.ppat.1005633

Brugger, S. D., Eslami, S. M., Pettigrew, M. M., Escapa, I. F., Henke, M. T., Kong, Y., et al. (2020). *Dolosigranulum pigrum* cooperation and competition in human nasal microbiota. *mSphere* 5:e0852-20. doi: 10.1128/mSphere.00852-820

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., et al. (2016). MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol.* 17:121. doi: 10.1186/s13059-016-0980-6

Byrd, A. L., and Segre, J. A. (2016). Adapting Koch's postulates. *Science* 351, 224–226. doi: 10.1126/science.aad6753

Cespedes, C., Saïd-Salim, B., Miller, M., Lo, S. H., Kreiswirth, B. N., Gordon, R. J., et al. (2005). The clonality of *Staphylococcus aureus* nasal carriage. *J. Infect. Dis.* 191, 444–452. doi: 10.1086/427240

Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* 31, 343–366. doi: 10.1146/annurev.ecolsys.31.1.343

De Vos, M. G. J., Zagorski, M., McNally, A., and Bollenbach, T. (2017). Interaction networks, ecological stability, and collective antibiotic tolerance in polymicrobial infections. *Proc. Natl. Acad. Sci. U.S.A.* 114, 10666–10671. doi: 10.1073/pnas.1713372114

Dedrick, S., Akbari, M. J., Dyckman, S., Zhao, N., Liu, Y.-Y., and Momeni, B. (2020). Impact of temporal pH fluctuations on the coexistence of nasal bacteria. *bioRxiv* [Preprint], doi: 10.1101/2020.09.15.298778

Eddison, J. C., and Ollason, J. G. (1978). Diversity in constant and fluctuating environments [11]. *Nature* 275, 309–310. doi: 10.1038/275309a0

Escapa, I. F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F. E., and Lemon, K. P. (2018). New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *mSystems* 3:e0187-18. doi: 10.1128/mSystems.00187-118

Fisher, C. K., and Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One* 9:e0102451. doi: 10.1371/journal.pone.0102451

Fukuyama, J., Rumker, L., Sankaran, K., Jeganathan, P., Dethlefsen, L., Relman, D. A., et al. (2017). Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput. Biol.* 13:e1005706. doi: 10.1371/journal.pcbi.1005706

Gerber, G. K. (2014). The dynamic microbiome. *FEBS Lett.* 588, 4131–4139. doi: 10.1016/j.febslet.2014.02.037

Gilbert, J. A., and Lynch, S. V. (2019). Community ecology as a framework for human microbiome research. *Nat. Med.* 25, 884–889. doi: 10.1038/s41591-019-0464-9

Goldford, J. E., Lu, N., Bajiæ, D., Estrela, S., Tikhonov, M., Sanchez-Gorostiaga, A., et al. (2018). Emergent simplicity in microbial community assembly. *Science* 361, 469–474. doi: 10.1126/science.aat1168

Grover, J. P. (1988). Dynamics of competition in a variable environment: experiments with two diatom species. *Ecology* 69, 408–417. doi: 10.2307/1940439

Hehar, S. S., Mason, J. D. T., Stephen, A. B., Washington, N., Jones, N. S., Jackson, S. J., et al. (2001). Twenty-four hour ambulatory nasal pH monitoring. *Clin. Otolaryngol. Allied Sci.* 24, 24–25. doi: 10.1046/j.1365-2273.1999.00190.x

Iwase, T., Uehara, Y., Shinji, H., Tajima, A., Seo, H., Takada, K., et al. (2010). *Staphylococcus epidermidis* Esp inhibits *Staphylococcus aureus* biofilm formation and nasal colonization. *Nature* 465, 346–349. doi: 10.1038/nature09074

Jiang, L., and Morin, P. J. (2007). Temperature fluctuation facilitates coexistence of competing species in experimental microbial communities. *J. Anim. Ecol.* 76, 660–668. doi: 10.1111/j.1365-2656.2007.01252.x

Johannessen, M., Sollid, J. E., and Hanssen, A.-M. (2012). Host- and microbe determinants that may influence the success of *S. aureus* colonization. *Front. Cell. Infect. Microbiol.* 2:56. doi: 10.3389/fcimb.2012.00056

Kaspar, U., Kriegeskorte, A., Schubert, T., Peters, G., Rudack, C., Pieper, D. H., et al. (2016). The culturome of the human nose habitats reveals individual bacterial fingerprint patterns. *Environ. Microbiol.* 18, 2130–2142. doi: 10.1111/1462-2920.12891

Kiryukhina, N. V., Melnikov, V. G., Suvorov, A. V., Morozova, Y. A., and Ilyin, V. K. (2013). Use of *Corynebacterium pseudodiphtheriticum* for elimination of *Staphylococcus aureus* from the nasal cavity in volunteers exposed to abnormal microclimate and altered gaseous environment. *Probiot. Antimicrob. Proteins* 5, 233–238. doi: 10.1007/s12602-013-9147-x

Kremer, C. T., and Klausmeier, C. A. (2013). Coexistence in a variable environment: eco-evolutionary perspectives. *J. Theor. Biol.* 339, 14–25. doi: 10.1016/j.jtbi.2013.05.005

Krismer, B., Liebeke, M., Janek, D., Nega, M., Rautenberg, M., Hornig, G., et al. (2014). Nutrient limitation governs *Staphylococcus aureus* metabolism and niche adaptation in the human nose. *PLoS Pathog.* 10:e1003862. doi: 10.1371/journal.ppat.1003862

Krismer, B., Weidenmaier, C., Zipperer, A., and Peschel, A. (2017). The commensal lifestyle of *Staphylococcus aureus* and its interactions with the nasal microbiota. *Nat. Rev. Microbiol.* 15, 675–687. doi: 10.1038/nrmicro.2017.104

Kurkjian, H. M., Akbari, M. J., and Momeni, B. (2020). The impact of interactions on invasion and colonization resistance in microbial communities. *bioRxiv* [Preprint], doi: 10.1101/2020.06.11.146571

Larsen, P. E., Field, D., and Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nat. Methods* 9, 621–625. doi: 10.1038/nmeth.1975

Man, W. H., de Steenhuijsen Piters, W. A. A., and Bogaert, D. (2017). The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat. Rev. Microbiol.* 15, 259–270. doi: 10.1038/nrmicro.2017.14

Margolis, E., Yates, A., and Levin, B. R. (2010). The ecology of nasal colonization of *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Staphylococcus aureus*: the role of competition and interactions with host's immune response. *BMC Microbiol.* 10:59. doi: 10.1186/1471-2180-10-59

Momeni, B., Chen, C.-C., Hillesland, K. L., Waite, A., and Shou, W. (2011). Using artificial systems to explore the ecology and evolution of symbioses. *Cell. Mol. Life Sci.* 68, 1353–1368. doi: 10.1007/s00018-011-0649-y

Momeni, B., Xie, L., and Shou, W. (2017). Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *eLife* 6:e25051. doi: 10.7554/eLife.25051

Mounier, J., Monnet, C., Vallaeys, T., Arditi, R., Sarthou, A.-S., Hélias, A., et al. (2008). Microbial interactions within a cheese microbial community. *Appl. Environ. Microbiol.* 74, 172–181. doi: 10.1128/AEM.01338-07

Niehaus, L., Boland, I., Liu, M., Chen, K., Fu, D., Henckel, C., et al. (2019). Microbial coexistence through chemical-mediated interactions. *Nat. Commun.* 10:2052. doi: 10.1038/s41467-019-10062-x

Proctor, D. M., and Relman, D. A. (2017). The landscape ecology and microbiota of the human nose, mouth, and throat. *Cell Host Microb.* 21, 421–432. doi: 10.1016/J.CHOM.2017.03.011

Ratzke, C., Denk, J., and Gore, J. (2018). Ecological suicide in microbes. *Nat. Ecol. Evol.* 2, 867–872. doi: 10.1038/s41559-018-0535-531

Ratzke, C., and Gore, J. (2018). Modifying and reacting to the environmental pH can drive bacterial interactions. *PLoS Biol.* 16:e2004248. doi: 10.1371/journal.pbio.2004248

Regev-Yochay, G., Raz, M., Dagan, R., Porat, N., Shainberg, B., Pinco, E., et al. (2004). Nasopharyngeal carriage of *Streptococcus pneumoniae* by adults and children in community and family settings. *Clin. Infect. Dis.* 38, 632–639. doi: 10.1086/381547

Relman, D. A. (2012). The human microbiome: ecosystem resilience and health. *Nutr. Rev.* 70, S2–S9. doi: 10.1111/j.1753-4887.2012.00489.x

Shenhav, L., Furman, O., Briscoe, L., Thompson, M., Silverman, J. D., Mizrahi, I., et al. (2019). Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLOS Comput. Biol.* 15:e1006960. doi: 10.1371/journal.pcbi.1006960

Shibasaki, S., Mobilia, M., and Mitri, S. (2020). Microbial species interactions determine community diversity in fluctuating environments. *bioRxiv* [Preprint], doi: 10.1101/2020.07.22.216010

Silverman, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S., and David, L. A. (2018a). Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* 6:202. doi: 10.1186/s40168-018-0584-3

Silverman, J., Shenhav, L., Halperin, E., Mukherjee, S., and David, L. A. (2018b). Statistical considerations in the design and analysis of longitudinal microbiome studies. *bioRxiv* doi: 10.11164/jjsps.27.3_633_1

Song, H.-S., Cannon, W., Beliaev, A., and Konopka, A. (2014). Mathematical modeling of microbial community dynamics: a methodological review. *Processes* 2, 711–752. doi: 10.3390/pr2040711

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., et al. (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol.* 9:e1003388. doi: 10.1371/journal.pcbi.1003388

Uehara, Y., Nakama, H., Agematsu, K., Uchida, M., Kawakami, Y., Abdul Fattah, A., et al. (2000). Bacterial interference among nasal inhabitants: eradication of *Staphylococcus aureus* from nasal cavities by artificial implantation of *Corynebacterium* sp. *J. Hosp. Infect.* 44, 127–133. doi: 10.1053/JHIN.1999.0680

Venturelli, O. S., Carr, A. V., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., et al. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.* 14:e8157. doi: 10.15252/MSB.20178157

Washington, N., Steele, R. J., Jackson, S., Bush, D., Mason, J., Gill, D., et al. (2000). Determination of baseline human nasal pH and the effect of intranasally administered buffers. *Int. J. Pharm.* 198, 139–146. doi: 10.1016/S0378-5173(99)00442-441

Wertheim, H. F., Vos, M. C., Ott, A., van Belkum, A., Voss, A., Kluytmans, J. A., et al. (2004). Risk and outcome of nosocomial *Staphylococcus aureus* bacteraemia in nasal carriers versus non-carriers. *Lancet* 364, 703–705. doi: 10.1016/S0140-6736(04)16897-16899

Yan, M., Pamp, S. J., Fukuyama, J., Hwang, P. H., Cho, D.-Y., Holmes, S., et al. (2013). Nasal microenvironments and interspecific interactions influence nasal microbiota complexity and *S. aureus* carriage. *Cell Host Microb.* 14, 631–640. doi: 10.1016/J.CHOM.2013.11.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 02 June 2021
doi: 10.3389/fmicb.2021.681485

Model Selection Reveals the Butyrate-Producing Gut Bacterium *Coprococcus eutactus* as Predictor for Language Development in 3-Year-Old Rural Ugandan Children

Remco Kort[1,2,3], Job Schlösser[1], Alan R. Vazquez[4†], Prudence Atukunda[5], Grace K. M. Muhoozi[5,6], Alex Paul Wacoo[1,2,7], Wilbert F. H. Sybesma[2], Ane C. Westerberg[8], Per Ole Iversen[5,9] and Eric D. Schoen[4]*

[1] Department of Molecular Cell Biology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [2] Yoba for Life Foundation, Amsterdam, Netherlands, [3] ARTIS-Micropia, Amsterdam, Netherlands, [4] Faculty of Bioscience Engineering, KU Leuven, Leuven, Belgium, [5] Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway, [6] Department of Human Nutrition and Home Economics, Kyambogo University, Kampala, Uganda, [7] Department of Medical Biochemistry, Makerere University, Kampala, Uganda, [8] Institute of Health Sciences, Kristiania University College, Oslo, Norway, [9] Department of Hematology, Oslo University Hospital, Oslo, Norway

Introduction: The metabolic activity of the gut microbiota plays a pivotal role in the gut-brain axis through the effects of bacterial metabolites on brain function and development. In this study we investigated the association of gut microbiota composition with language development of 3-year-old rural Ugandan children.

Methods: We studied the language ability in 139 children of 36 months in our controlled maternal education intervention trial to stimulate children's growth and development. The dataset includes 1170 potential predictors, including anthropometric and cognitive parameters at 24 months, 542 composition parameters of the children's gut microbiota at 24 months and 621 of these parameters at 36 months. We applied a novel computationally efficient version of the all-subsets regression methodology and identified predictors of language ability of 36-months-old children scored according to the Bayley Scales of Infant and Toddler Development (BSID-III).

Results: The best three-term model, selected from more than 266 million models, includes the predictors *Coprococcus eutactus* at 24 months of age, *Bifidobacterium* at 36 months of age, and language development at 24 months. The top 20 four-term models, selected from more than 77 billion models, consistently include *C. eutactus* abundance at 24 months, while 14 of these models include the other two predictors as well. Mann–Whitney *U* tests suggest that the abundance of gut bacteria in language non-impaired children (n = 78) differs from that in language impaired children (n = 61). While anaerobic butyrate-producers, including *C. eutactus*, *Faecalibacterium prausnitzii*, *Holdemanella biformis*, *Roseburia hominis* are less abundant, facultative anaerobic bacteria, including *Granulicatella elegans*, *Escherichia/Shigella* and *Campylobacter coli*, are more abundant in language impaired children. The overall predominance of oxygen

OPEN ACCESS

Edited by:
Dong-Yup Lee, Sungkyunkwan University, South Korea

Reviewed by:
*Sangyong Jung, Institute of Molecular and Cell Biology (A*STAR), Singapore*
Carmen Wacher, National Autonomous University of Mexico, Mexico

***Correspondence:**
Remco Kort
r.kort@vu.nl

†Present address:
Alan R. Vazquez, Department of Statistics, University of California, Los Angeles, Los Angeles, CA, United States

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *16 March 2021*
Accepted: *10 May 2021*
Published: *02 June 2021*

Citation:
Kort R, Schlösser J, Vazquez AR, Atukunda P, Muhoozi GKM, Wacoo AP, Sybesma WFH, Westerberg AC, Iversen PO and Schoen ED (2021) Model Selection Reveals the Butyrate-Producing Gut Bacterium Coprococcus eutactus as Predictor for Language Development in 3-Year-Old Rural Ugandan Children. Front. Microbiol. 12:681485. doi: 10.3389/fmicb.2021.681485

tolerant species in the gut microbiota was slightly higher in the language impaired group than in the non-impaired group ($P = 0.09$).

Conclusion: Application of the all-subsets regression methodology to microbiota data established a correlation between the relative abundance of the anaerobic butyrate-producing gut bacterium *C. eutactus* and language development in Ugandan children. We propose that the gut redox potential and the overall bacterial butyrate-producing capacity in the gut are important factors for language development.

Keywords: gut-brain-axis, butyrate, *Coprococcus eutactus*, language development, all subsets regression, mixed integer optimization, metagenomic aerotolerant predominance index

INTRODUCTION

There is an accumulating amount of evidence for a role of the gut microbiota in brain function and development via the so-called microbiota-gut-brain-axis, as recently reviewed by Cryan et al. (2019). The communication along this axis is bidirectional. Communication from the brain to the gut occurs through signals to change bowel movements and intestinal permeability, which in turn changes the enteric microbiota composition, its metabolic activity and response signal. The gut microbiota signals to the brain via stimulation of intestinal host immune cells, eliciting a cytokine response. In addition, signals are transferred to the brain through bacterial metabolites, including short chain fatty acids (SCFA's). This results in altered neurotransmitter release, hormone secretion and induction of vagus nerve signaling to the brain (Rhee et al., 2009; Bienenstock et al., 2015; Jameson et al., 2020).

In this study we investigated the correlation between gut microbiota composition and language ability of 3-year-old rural Ugandan children, as assessed by the Bayley Scales of Infant and Toddler Development (BSID-III) composite scores for language development (Albers and Grieve, 2007). The scales provide comprehensive development measures for children up to 42 months and have been adapted for appropriate use among children in rural Uganda (Muhoozi et al., 2016). The data used in this study were collected during a follow-up trial of a two-armed, open cluster-randomized education intervention regarding nutrition, child stimulation and hygiene among mothers of children in the Kisoro and Kabale districts of South-West Uganda (Muhoozi et al., 2018). The intervention did not lead to any significant changes in the gut microbiota diversity compared with the control group at phylum or genus level. Neither did we observe any significant differences between the two study groups in the Shannon diversity index at 20–24 and 36 months, respectively. However, the Shannon diversity index of the gut microbiota increased significantly in both study groups from 24 to 36 months (Atukunda et al., 2019). Further analysis of the changes associated with the gut microbiota in the transition from 24 to 36 months revealed that there was a notable shift from autochthonous (endogenous) to allochthonous (plant-derived) *Lactobacillus* species, and a correlation of *Lactobacillus* with stunting, most probably resulting from the change in the children's diet from breast milk to solid, plant-based foods (Wacoo et al., 2020). As follow-up to this study we further investigate here correlations between the gut microbiota of these children with language development.

It should be noted that predictors for current cognition parameters in children may not only be found in past values of these parameters, but also in current and past gut microbiota compositions. This is supported by longitudinal studies that indicate a maturation program of the human gut microbiome in the first 3 years of life, consisting of distinct phases of microbiome progression (Backhed et al., 2015; Stewart et al., 2018). Suitable predictors are usually found by fitting models including the predictors being assessed and comparing the fit of the model with the fit of a model that does not include these predictors. This poses a non-trivial problem, because the number of different models that can be fitted grows exponentially with the number of potential predictors, so it is not feasible to fit all possible models and compare their fit. In addition, the predictors can be correlated so that different sets of predictors can explain the response variable of interest equally well. In the present paper, we successfully address the above mentioned problems in data analysis of the gut microbiota from rural Ugandan children. Our key finding is that abundance of butyrate-producing bacterium *Coprococcus eutactus* in the gut microbiota at 24 months predicts language development in these children at 36 months.

MATERIALS AND METHODS

Study Design and Data Collection

The data used in this study were collected during a follow-up trial of a two-armed, open cluster-randomized education intervention regarding nutrition, stimulation and hygiene among impoverished mothers of children in the Kisoro and Kabale districts of South-Western Uganda (Muhoozi et al., 2018). The purpose of the study by Muhoozi et al. was to assess the effects of a nutrition education intervention, delivered in group meetings to impoverished mothers, on child growth, cognitive development and gut microbiota in rural Uganda. Developmental outcomes were assessed with the BSID-III composite scores for cognitive (primary endpoint), language and motor development. Other outcomes included gut microbiota compositions.

Stool samples were collected from 139 children at the age of 20–24 months and at 36 months and shipped to Netherlands for DNA extraction (Atukunda et al., 2019). Quantitative PCR was performed to determine the relative amount of bacterial template

and amplicon sequencing was carried out as previously described (de Boer et al., 2015; Parker et al., 2018). In summary, V4 16S rRNA gene amplicon sequencing was carried out by paired end sequencing conducted on an Illumina MiSeq platform (Illumina, Netherlands). Taxonomic names were assigned to all sequences using the Ribosomal Database Project (RDP) naïve Bayesian classifier with a confidence threshold of 60% (Wang et al., 2007) and the mothur-formatted version of the RDP training set v.9 (Schloss et al., 2009). All 16S rRNA amplicon paired end reads of the gut microbiota samples sequenced in this study are accessible at BioProject PRJNA517509 (Kort, 2019).

Language development was determined by the Bayley Scales of Infant and Toddler Development 3rd edition (BSID−III) using the language subscale. The BSID-III provides comprehensive development measures with children up to 42 months and has been adapted for appropriate use among children in rural Uganda (Muhoozi et al., 2016, 2018). The BSID-III language component focuses on pre-linguistic behaviors, communication and social routines in addition to expressive and receptive language skills. The children's performance was scored according to the guidelines in the administration manual and the raw scores from expressive and receptive subscales were summed up and converted to composite scores using BSID−III conversion tables. In the reference material of United States children the mean score after conversion is 100.

Model Selection Using Mixed Integer Optimization

Model selection strategies should reveal sets of predictors that explain the data equally well, if such is the case. Best subset selection (Miller, 2002) based on Ordinary Least Squares (OLS) returns the best k models with p predictors each, so that the common predictors in the best models form a solid basis to explain the response variable of interest and the predictors that differ among the best models point to alternative interpretations to explain the same variable. However, until recently, subset selection could only be performed when the total number of predictors t is fairly small, say, $t < 30$. Therefore, best subset selection used to be a less attractive model selection technique for research that assesses many parameters. Obviously, one could perform OLS-based forward selection to select predictors (Miller, 2002). This approach has the disadvantage that the resulting models comprise a single path in multidimensional space. That is, there is one model for each number of predictors up to p. There is no guarantee that the model with p predictors corresponds with the model of the same size from best subset selection.

Bertsimas et al. (2016) proposed methodology to select the best model with p out of t predictors with t in the 100 s. Their approach is based on Mixed Integer Optimization (MIO). The key innovation is that searching unpromising sets of predictors is cut off in an early stage of the calculations so that not all of the models with p predictors have to be assessed. In the original form, just one model with p predictors is returned along a range of values for p extended the original form to obtain the second-best up to k-th best models of given size as well (Vazquez et al., 2020). The method thus results in a list of models compatible with the data. The authors further employ a powerful visualization method to reveal possible alternative ways to explain the same variable. For example, one might observe that either the effect of predictor X or the effect of predictor Y is in the best ten models that link language development to four predictors, but the models do not include both of them.

For ease of reference, we call the method of Vazquez et al. (2020) MIO after its core element. It was developed primarily with applications in statistical design of industrial experiments in mind. The data in these cases usually have few observations and many controllable experimental factors. This is similar to field studies on human microbiota compositions where the number of cases is much smaller than the number of species.

A key element of MIO is best-subset selection, which finds the best fitting model with p parameters as measured by the model's residual sum of squares. Current state-of-the-art algorithms for best-subset selection, as implemented in SAS 9.4 or JMP 14, or in the "leaps" package in R, which is based on (Furnival and Wilson, 1974), do not allow solving the problem when the search is over more than $t = 30$ predictors (Vazquez et al., 2020). Bertsimas et al. (2016) proposed a formulation for the best subset selection in terms of a MIO problem. Modern optimization solvers such as (Gurobi,, 2017), do permit searching over a large number of potential predictors. The goal function to be minimized is

$$\min_{\hat{\beta},\hat{\eta},\, z} \hat{\eta}^T\hat{\eta} - 2(X^T y)^T + y^T y \tag{1}$$

In this equation, $\hat{\eta}$ is an $N \times 1$ vector of fitted values, $\hat{\beta}$ is a $t \times 1$ vector of coefficients for the regression equation, y is the $N \times 1$ vector of observations, X is an $N \times t$ matrix of predictors, and z is a $t \times 1$ indicator vector that indicates whether or not the corresponding elements of $\hat{\beta}$ are non-zero. The goal function (1) is a version of the residual sum of squares rewritten to reduce the number of quadratic variables from t to N. This is useful because in our application there are many more potential predictors than there are subjects.

An optimization model allows for the minimization of the goal function under constraints. The constraints proposed by Bertsimas et al. (2016) are:

$$z_u \in \{0, 1\}, u = 1, \cdots, t \tag{2}$$

$$(1 - z_u)\,\hat{\beta}_u = 0,\ \ u = 1, \cdots, t \tag{3}$$

$$\sum_{u=1}^{t} z_u \le p \tag{4}$$

$$\hat{\eta} = X\hat{\beta} \tag{5}$$

Constraint (2) defines the individual elements z_u of the vector z as binary variables. Constraint (3) features the regression coefficients for the individual predictors $\hat{\beta}_u$. The constraint specifies that $\hat{\beta}_u$ can be non-zero if z_u equals 1 and that $\hat{\beta}_u$ is exactly zero if z_u equals 0. Constraint (4) restricts the regression model to at most p non-zero parameters. Finally, constraint (5) defines $\hat{\eta}$ as the fitted values matching the coefficients in $\hat{\beta}$. The model (1)–(5) returns for each value of p specified by the data

analyst the best fitting model as measured by the residual sum of squares. Vazquez et al. (2020) extended the application potential by proposing further constraints to obtain the second best, third best etc. model for each value of p. For example, if parameters 1 and 2 define the best fitting model with $p = 2$ terms, the constraint $z_1+z_2 < 2$ is added to the constraints (2)–(5) and the model is rerun. The constraints prevent simultaneous inclusion of parameters 1 and 2 in the new model so that a second best model results. Vazquez et al. (2020) implemented the MIO model in Python using (Gurobi,, 2017) as the solver of the optimization and used the raster plots of Wolters and Bingham (2011) to visualize the models. For this purpose, the predictors are rescaled so that they all have the same length. The raster plot represents each model with p parameters as p pixels that are darker or lighter according to the size of the respective coefficients. Each predictor has its own horizontal coordinate and each model has its own vertical coordinate. The models are ordered according to the number of non-zero coefficients and, subsequently, their residual sum of squares. Predictors that often occur in the models form a band in the plot.

Promising predictors of the language development of 139 children at 36 months of age were selected for the MIO approach described above. The data included a total of 1170 potential predictors (**Supplementary File 1**), including one parameter indicating whether or not the mother of the child was included in the education intervention group), six anthropometric and cognitive parameters when the children were 24 months 542 gut microbiota composition related parameters at 24 months and 621 parameters at 36 months. Subsequently, the 20 best models were established with 1–4 predictors in terms of their residual standard deviation. The best 4-term models were selected from more than 77 billion models, which is the number of ways one can choose four objects out of 1170.

In order to compare the results obtained by the MIO approach to those obtained by a conventional statistical method, the same data were also evaluated by the Mann–Whitney U test (Mann and Whitney, 1947). The statistical distribution of the abundance data is a mixture of binary date (absence or presence of the species) and rational data (measure of abundance if present). Therefore, we used the non-parametric Mann–Whitney test rather than a parametric alternative. Using the Mann–Whitney test we investigated which bacterial species had a different abundance in the gut microbiota of children that scored equal or above average for language development when compared with those that scored lower than average. For this purpose, all the 139 children at the age of 3 years old were divided into a "language impaired" or "language below average" group with a BSD-III score below the mean value of 100 ($n = 61$), and a "language non-impaired" or "equal or above average" group with a BSID-III score of 100 or higher ($n = 78$).

PCR-Based Identification of *Coprococcus eutactus* in Stool Samples

For the experimental identification of *C. eutactus* in stool samples, species-specific primers were designed for the 16S rRNA gene via primer-BLAST™ (Ye et al., 2012): forward-primer 785F 5′-GGGTTCCAAAGGGACTCGG-3′ and reverse primer 1412R 5′-CAGCTCCCTCTTGCGGTT-3′. The oligonucleotides were manufactured by Biolegio™ (Nijmegen, Netherlands) and delivered in 100 μL TAE-buffer with a concentration of 100 μM. DNA was released from the stool samples in nuclease-free milliQ by heating an Eppendorf tube at 95°C for 10 min. The PCR mix contained 12.5 μL GoTaq™ mastermix, 2.5 μL of 10 μM forward primer, 2.5 μL of 10 μM reverse-primer, 5.0 μL nuclease-free milliQ, 2.5 ul template DNA. PCR-samples were placed in the PCR machine (Biometra™, model *Tgradient*) with 1 cycle of 95°C for 5 min; 30 cycles of 95°C for 30 s, 60°C for 30 s and 72°C for 1 min, completed 72°C for 5 min. Products were analyzed by the use of a 1.5% agarose gel with ethidium bromide in TAE-buffer. The PCR was validated by the use of genomic DNA from the cultivated *C. eutactus* type strain ATCC 27759 as a positive control. This strain was obtained from the German Collection of Microorganisms and Cell Cultures (DSM strain number 107541) and cultivated under anaerobic conditions in chopped meat casitone (CMC) medium as described by the supplier.

The Core Microbiota of Ugandan Children at 24 and 36 Months of Age

For the definition of the core, the bacterial 16S rRNA gene amplicon sequencing dataset of the Ugandan children's feces cohort (139 subjects, measured at 24 months) was used, obtained from the study of Atukunda et al. (2019). The 1163 bacterial V4-region 16S rRNA gene sequences, delivered in Microsoft™ Excel format, were annotated to bacterial genus and species via the local BLAST in CLC Workbench™ Version 20.0 computational software. All sequence abundances were grouped to species-level and ordered by most prevalent to least prevalent. Bacterial V4-region 16S rRNA amplicon sequences with hits of more than three species with identical identity-scores were grouped to genus level (e.g., *Bifidobacterium*). The core was composed by the top 50 most prevalent bacterial species at a 0.1% relative abundance detection threshold, as described previously (Shetty et al., 2017). From the composed core a heat map was created by MeV™ software (Saeed et al., 2003), thereby including a 0.1–100 logarithmic scale for the x-axis threshold at percentage of relative abundance and representing the prevalence via color-scaling for each of the 50 relative abundance detection thresholds.

Assessment of Metagenomic Aerotolerant Predominance Index (MAPI)

To assess aerobic/anaerobic balance in the gut microbiota samples of our cohort we used the Metagenomic Aerotolerant Predominance Index (MAPI) (Million and Raoult, 2018), based on a previously published database with a list of bacteria and their aerotolerant or obligate anaerobic metabolism (Million et al., 2016). This MAPI index indicates the ratio of the metagenomic relative abundance of aerotolerant species and the relative abundance of strict anaerobes. From the taxonomic assignment of amplicon sequence variants (ASV's) of each of the 139 stool bacterial communities of Ugandan children (**Supplementary File 1**), we calculated the total number of reads that corresponded to strict aerotolerant or anaerobic

bacteria. We then calculated the ratio of aerotolerant relative abundance to strict anaerobic relative abundance. This ratio was >1 for aerotolerant predominance and <1 for strict anaerobic predominance. In order to fit a lognormal distribution, the natural logarithm of the aerotolerant ratio was calculated for each metagenome for further analysis. The MAPI corresponds to the variable "ln(Ae/Ana)."

Ethical Approval

All mothers gave written or thumb-printed, informed consent to participate and could decline an interview or assessment at any time. The study was approved by The AIDS Support Organization Research Ethics Committee (no. TASOREC/06/15-UG-REC-009) and by the Uganda National Council for Science and Technology (no. UNCST HS 1809) as well as by the Norwegian Regional Committee for Medical and Health Research Ethics (no. 2013/1833). The trial was registered at clinicaltrials.gov (NCT02098031).

RESULTS

Abundance of *Coprococcus eutactus* Is a Predictor for Language Development

The application of the MIO approach to identify predictors for language development in Ugandan children at 36 months of age resulted in the raster plot of **Figure 1A**. For this visualization, we normalized the predictors and the language development score such that their means are zero and their standard deviations are one. A coefficient thus expresses the increase in the response, in terms of multiples of its standard deviation, if a predictor is increased by one standard deviation. As co-occurrences can only be recorded in models with two or more parameters, we ignore models with a single parameter in our evaluation. The red vertical band with horizontal axis label 5 shows that, with a few exceptions, the best models with 2–4 parameters include the language development of the children at 24 months and that its coefficient is positive. This parameter is included in 52 of the 60 models with 2–4 parameters. The figure shows that in 7 of the 8 remaining cases, cognition at 24 months (horizontal axis label 4) replaces language development at 24 months in the model. The MIO methodology shows here alternative explanations of the same data by correlated predictors. Indeed, the Pearson correlation coefficient of the language ability and cognition parameters equals 0.7. In spite of this correlation, the much higher frequency of occurrence of the language ability at 24 months suggests that this parameter should be included in favor of cognition at 24 months. Further red bands can be observed at horizontal axis labels 281 and 563, respectively. These bands correspond with relative abundances of *C. eutactus*, and *Bifidobacterium* from the gut microbiota at 24 and 36 months of age, respectively. The abundance of *C. eutactus* occurs in 42 of the 60 models with 2–4 parameters, while the abundance of *Bifidobacterium* occurs in 19 of these models. A total of 18 of the models include both parameters. Species identities were verified with the BLAST tool. They led to a species assignment on the basis of a 100% identity match with the partial 16S rRNA sequence of *C. eutactus* strain ATCC 27759 (Holdeman and Moore, 1974) over the total length of the sequenced V4 region of 253 base pairs. The assignment of the species *C. eutactus* is unambiguous, but sequences of *Bifidobacterium longum* and *Bifidobacterium breve* are both aligning to the 16S rRNA V4 sequence with a match of 100% identity, therefore we refer to parameter 563 as a match to the *B. longum* group (see also below).

Figure 1 further shows that some predictors enter the fitted models occasionally. The most frequently occurring parameter after the *B. longum* group is *Intestinibacter bartlettii* (previously known as *Clostridium bartlettii*) at 24 months of age (horizontal label 348). This identification was based on a unique 100% identity match of the V4 amplicon with the partial 16S rRNA gene sequence of the type strain *I. bartlettii* strain WAL 16138 (Song et al., 2004). As this parameter enters only in 8 out of the 60 models, there is no powerful evidence that it should be included in a regression model. The residual standard deviations for the 80 models in **Figure 1A** were plotted against the number of predictors in the models in **Figure 1B**. The latter figure shows that, for four predictors, many of the 20 subsets explain the data equally well, so this application of a method to reveal the common elements in these subsets is warranted. Exploration of the common elements points to a three-parameter model with parameters five (language development of the children at 24 months), 281 (abundance of *C. eutactus*) and 563 (abundance of the *B. longum* group at 36 months), respectively. This model turns out to be the best three-parameter model. The figure shows that its residual standard deviation clearly stands out from the remaining 19 models. We conclude that a model including these three parameters explains these data best. In addition, we checked the results of MIO with the gut microbiota data expressed on the genus level comprising a set of 293 potential predictors. The results show that the genus *Coprococcus* is present in five of the top ten four-term models confirming its importance as a predictor for language development (**Supplementary File 2**).

The linear regression model for BSID-III language development at 36 months with all parameters on their original scale is summarized in **Table 1**. For the intercept and each predictor in the model, the entries in columns three to six show the regression coefficient, its standard error, the ratio of the coefficient to its standard error (t Ratio) and the *P*-value of this ratio. The residual standard deviation of the model is 12 based om 135 degrees of freedom. This measure quantifies the variability unexplained by the model. The model's *F* value equals 21.5. This measure indicates how much larger the model mean square is when compared to the unexplained variance. The adjusted R^2 value of 31% is the percentage of variation accounted for by the model, adjusted for the number of parameters.

All the model coefficients are positive so that higher values of previous BSID-III language development, previous abundance of *C. eutactus* and current abundance of *B. longum* point to higher language development at 36 months. The large values of the abundance coefficients can be explained by the measurement scale. The observed relative abundances of *C. eutactus* at 24 months (ID 281) and *B. longum* at 36 months of age (ID 563) are at most 4.5%. The model in **Table 1** includes the language development score recorded when the children were at the age

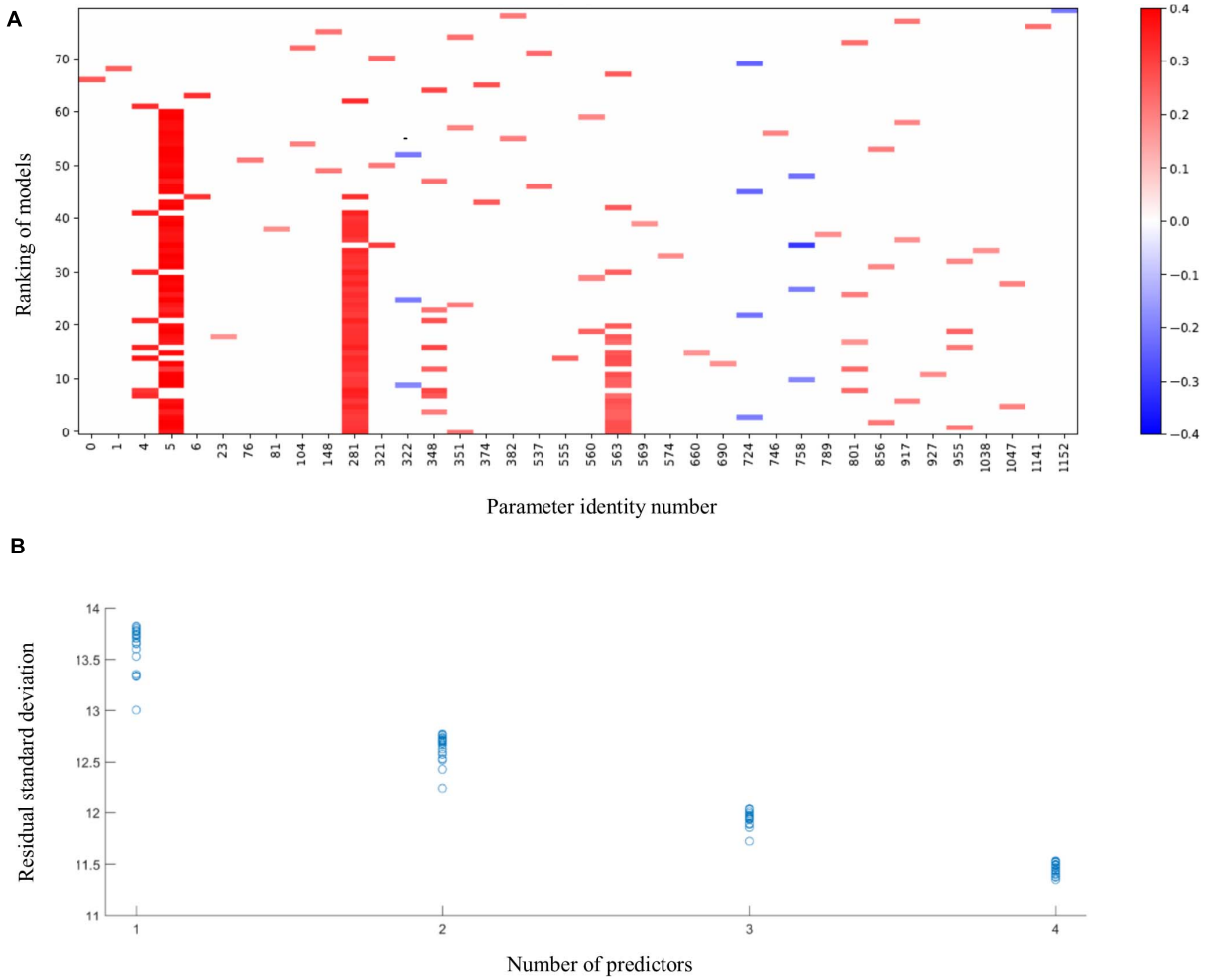


FIGURE 1 | The 20 best fitting models of 1 up to 4 terms linking language development of children aged 36 months to membership of the intervention or control group, values of 6 developmental parameters of the same children when aged 20–24 months, their microbiota composition at that age, and microbiota composition at 36 months. **(A)** Predictors in the models. Language development and all predictors are normalized to a mean of 0 and a standard deviation of 1. Horizontal labels correspond with the predictors' identification number in the data file. The parameters include the following scores at 24 months: the intervention indicator (0); height to age, HAZ (1); weight to age, WAZ (2); weight to height, WHZ (3); cognition (4); language development (5); motoric development (6). These are followed by the gut microbiota parameters at 24 months (parameters 7–548) and at 36 months (parameters 548–1169). Vertical coordinates 1–20, 21–40, 41–60, and 61–80 show best fitting models with 4, 3, 2, and 1 terms, respectively. Blue pixels correspond with negative coefficients and red pixels correspond with positive coefficients. Intensity of the pixels increases with size of the model coefficients. **(B)** Residual standard deviations plotted against the number of terms in the models for language development; language development in original units.

TABLE 1 | Prediction model for language development.

ID	Parameter	Coefficient	Standard error	*t* Ratio	*P*-value	Cross-validation
	Intercept	57	7.8			
5	language (24 months)	0.44	0.082	5.4	<0.001	0.44 ± 0.036
281	*Coprococcus eutactus* (24 months)	1929	430	4.5	<0.001	1927 ± 178
563	*Bifidobacterium longum* group (36 months)	417	114	3.7	<0.001	417 ± 32

Coefficients were calculated for the BSID-III scores of children aged 36 months by the MIO approach as indicated in ***Figure 1****.*
ID, identification number in data file.
Residual standard deviation = 12; degrees of freedom = 135; model F value = 21.5; Adjusted R^2 (%) = 31. Cross-validation: data was split in six sets of 20 children and one set of 19 children. Coefficients were calculated leaving out each of the seven sets in turn. The table shows means ± standard deviations of the seven coefficient estimates.

of 24 months. The interpretation of this finding is that children that had the same development score when they were of that age, differ in their subsequent language development according to their microbiota composition at that age and the composition at their current age.

The number of potential predictors far surpasses the number of children tested leading to the possibility of happenstance correlations between language development scores and predictor variables. Therefore, we cross-validated the model by random division of the data into six subsets of 20 children and one subset of 19 children. Subsequently, we fitted the model with language (24 month), *C. eutactus* (24 month), and *B. longum* (36 month) on the data leaving out each of the subsets in turn, and predicted the language score (36 month) of the children in the left out subset. The model residuals have been indicated in **Figure 2**. The cross validation root mean square error averaged over the seven subsets was 12.1. The residual standard deviation of the model fitted on the entire dataset was 12.0. This value is only slightly smaller than the average cross-validation root mean square error. The standard deviation of the language score was 14.4. The models did not overfit the data, because there is still a substantial unexplained variation. However, this unexplained variation should not be addressed by the addition of more terms in the model. The averages and standard deviations of the coefficients over the seven fitted models are shown in column seven of **Table 1**.

Increased Prevalence of *Coprococcus eutactus* in the Core Gut Microbiota Over Time

In order to evaluate the prevalence of *C. eutactus* in the microbiota of children of 24 and 36 months in relation to other highly abundant members of the intestinal microbiota, we carried out a comparative analysis of the core gut microbiota from the Ugandan children at 24 and 36 months (**Figure 3**). It should be noted that for this analysis all ASV's assigned to the same species have been pooled together. Both cores appear to be rather comparable in composition (80% of the species are present in both cores). However, at 24 months only the top three species, including *Faecalibacterium prausnitzii*, *Prevotella copri*, and *Blautia wexlerea*, were highly prevalent (>90%, detection threshold 0.1%), while a set of ten species is highly prevalent among Ugandan children at 36 months, in line with a decrease in the variation in the gut microbiota composition among children at higher age. The overall prevalence and abundance of *Bifidobacterium* species increased at 36 months compared to 24 months (from position 24 to 10), although this is not the case for ASV's matching to *B. longum* (see ASV ID 563 in **Figure 4**), in agreement with the notion that the relative abundance of this species reduces when children are no longer breast fed. The butyrate producing species *F. prausnitzii* was the most prevalent bacterial species in the datasets of both ages and is present in all Ugandan children in our study at the age of 36 months. Noteworthy, both microbiota cores clearly show a typical *Prevotella* gut microbiota type, in agreement with previous observations that led to the assignment of *Bacteroides* and *Prevotella* as biomarkers of diet and lifestyle in Western and non-Western subjects, respectively (Gorvitovskaia et al., 2016), and references herein. Accordingly, *Prevotella* species, such as *P. copri*, show much higher relative abundance among the majority of subjects in both heat maps of 24 and 36 months than *Bacteroides* species, represented in the core only by *Bacteroides xylanolyticus*. The gut bacterium *C. eutactus* is also represented in both cores, be it at relatively low prevalence and abundance levels; species position 44 at the age of 24 months and position 37 at the age of 36 months. The prevalence of the *C. eutactus* ASV ID 281 among the Ugandan children in this study increased from 24 to 36 months from 62 to 81%, although the average relative abundance was slightly lower (**Figure 4**). This increased prevalence over time was also evident in the core gut microbiota; all 11 ASV's matching to *C. eutactus* showed an increase from 38 to 44% at 0.1% abundance threshold (**Figure 3**). These observations and our best fitting model are in agreement with the notion that early acquirement of *C. eutactus* (before or at 24 months) is a beneficial factor for language development.

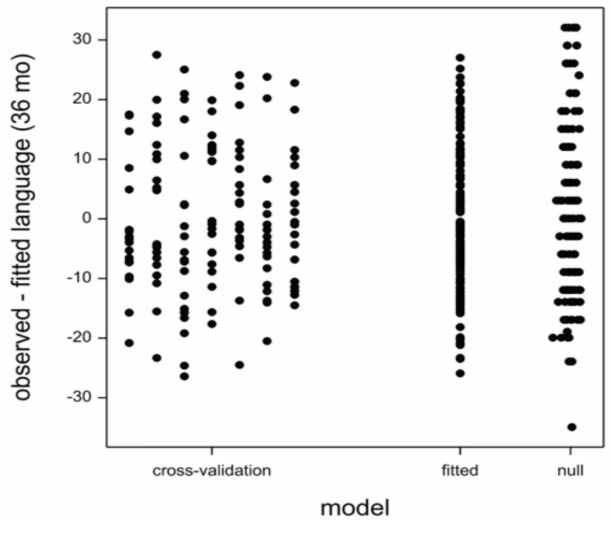


FIGURE 2 | Model: residuals from Cross-validation: residuals for left-out sets in cross validation; Root mean square error (RMSE) = 12.1. Fitted: residuals from the model fitted to all the data; RMSE = 12.0. Null: residuals form a model with only an intercept; RMSE = 14.4.

Butyrate-Producing Species More Abundant in Children With Above Average Language Development

The presence of the butyrate-producing bacterium *C. eutactus* was confirmed in the fecal samples of the Ugandan children in our cohort by PCR using specific primers designed in this study for *C. eutactus*. PCR-analysis of a random set of stool samples from the rural Ugandan children showed a product in 75% of the fecal samples in line with the prevalence range (62–81%) found for *C. eutactus* in our 16S rRNA gene sequence data from the children's stool samples. In order to further substantiate the results obtained by the MIO approach, we also checked with

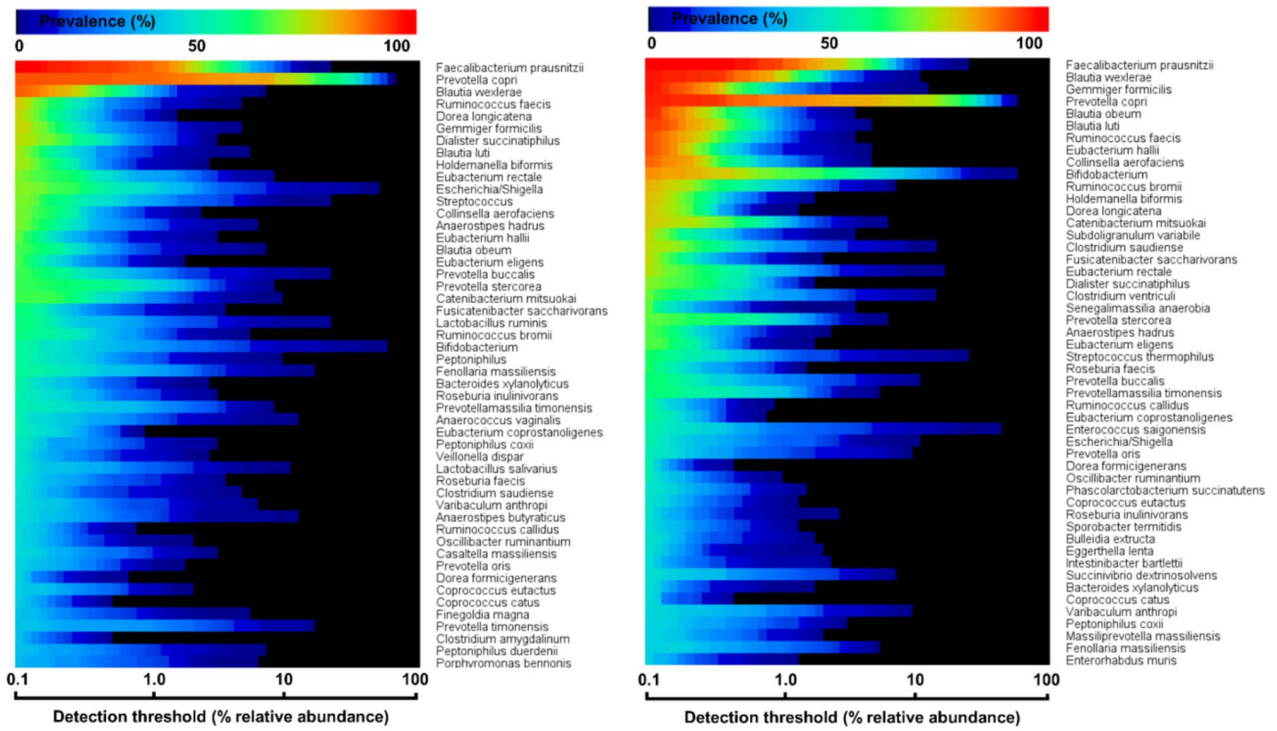


FIGURE 3 | Heat maps of top 50 most prevalent bacteria in the fecal microbiota of Ugandan children at the age of 24 months **(A)** and 36 months **(B)**, as determined by 16S rRNA gene amplicon sequencing. The color gradient indicates the prevalence (see top-legend) at the detection threshold of the relative abundance (%) presented at the *x*-axis with a logarithmic scale. The *y*-axis indicates the order of most prevalent bacteria at a detection threshold of 0.1% abundance. Unambiguous species assignments include *Dialister succinatiphilu*, *Dialister propionicifaciens*; *Lactobacillus salivarius*, *Lactobacillus ruminis*; *Clostridium saudiense*, *Clostridium disporicum*; *Varibaculum anthropi*, *Varibaculum cambriense*; *Prevotella oris*, *Prevotella albensis*, *Prevotella salivae*; *Clostridium amygdalinum*, *Clostridium methoxynbenzovorans*.

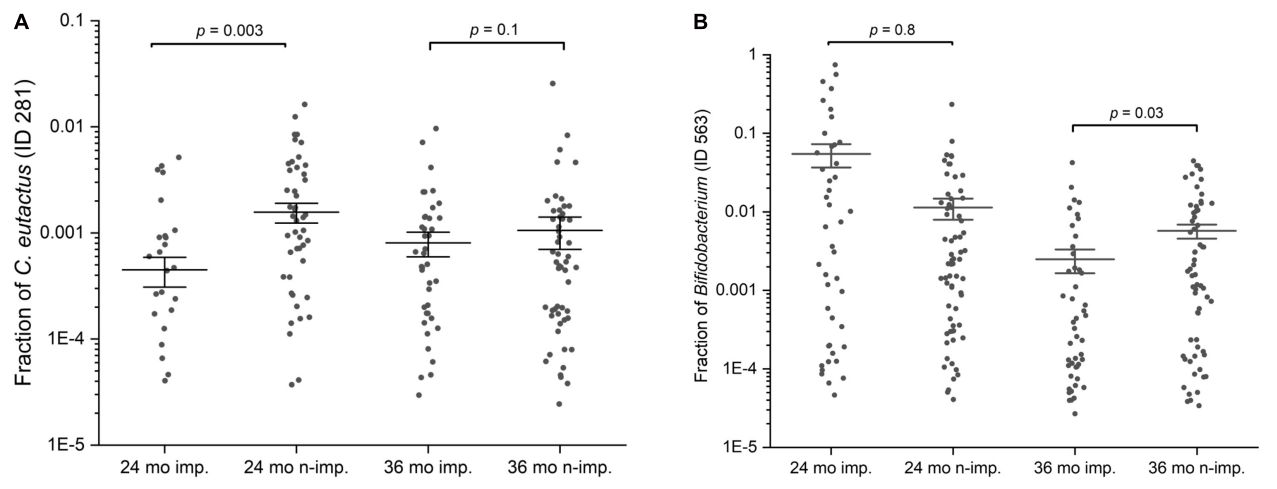


FIGURE 4 | Scatter interval plots of the fraction or relative abundance of *C. eutactus* and *Bifidobacterium*. **(A)** Fraction of *Coprococcus eutactus* (amplicon sequence variant ID 281) in the gut microbiota of 139 Ugandan children at the age of 24 and 36 months, **(B)** Fraction of *Bifidobacterium longum* group (amplicon sequence variant ID 563) in the gut microbiota of Uganda children aged 24 and 36 months. *P*-values were calculated with the two sided Mann–Whitney *U* test for language impaired ($n = 61$) and language non-impaired groups ($n = 78$) of the children.

a conventional statistical method (the Mann–Whitney *U* test) which bacterial species had a different abundance in children that scored equal or above average for language development when compared with children that scored below average. We first checked for the presence of the specific ASV's predicting language development by the MIO approach. We found that the relative abundances of the identified ASV's in our best fitting model of *C. eutactus* at 24 months (ID 281) and *Bifidobacterium* at 36 months (ID 563) were significantly different in both groups according to the two sided Mann–Whitney *U* test, with *P*-values of 0.003 and 0.03, as presented in **Figure 4**.

Out of the 542 gut microbiota ASV parameters at 24 months, 397 matched to a bacterial species with an identity score of 97% or higher. Using the latter composition parameters, we employed the two sided Mann–Whitney *U* tests to explore on a per species basis differences in abundance of these parameters between 3-year old children that scored equal or above average for language development and children that scored below average. If these two language groups would not differ in microbiota composition, we would expect that 20 out of the 397 tests have *P*-values below 0.05. Instead, twenty-five of these tests had such a *P*-value. **Table 2** lists the corresponding ASV's. Nineteen ASV's were more abundant in the equal or above average group and six ASV's were more present in the below average group. Among these were also other unique sequences matching to parameters identified in the predictive models as presented in **Figure 1**, including *Bifidobacterium* (ID 23), and *I. bartlettii* (ID 348).

A number of other striking features emerge from this Mann–Whitney *U* test. The list in **Table 2** contains five unique 16S rRNA gene ASVs which show a match with the genus *Bifidobacterium*. Although the V4 16S rRNA gene amplicon sequence does not allow for unambiguous assignment of species for this genus, we can assign these ASV's to three distinct *Bifidobacterium* species groups: the *Bifidobacterium catenelatum*, *adolescentis* and *longum* groups (see **Table 2**). Members of the first two groups show at 24 months a positive correlation with

TABLE 2 | Two-tailed Mann–Whitney *U* test for relative abundance of bacterial species equal or above average and below average language ability groups.

ID	Bacterial species	Identity (%)	*P*-value	core member	obligate anaerobic
Species with higher relative abundance in non-impaired language group					
281	*Coprococcus eutactus*	100.0	0.003	yes	yes
23	*Bifidobacterium catenulatum*[1]	99.6	0.003	yes	yes
20	*Bifidobacterium adolescentis*[2]	99.6	0.005	yes	yes
357	*Faecalibacterium prausnitzii*	97.6	0.006	yes	yes
14	*Bifidobacterium adolescentis*[2]	100.0	0.008	yes	yes
406	*Holdemanella biformis*	97.6	0.010	yes	yes
294	*Roseburia hominis*	100.0	0.010	no	yes
316	*Eubacterium eligens*	99.2	0.013	yes	yes
513	*Campylobacter troglodytis*	98.8	0.015	no	no
25	*Bifidobacterium adolescentis*[2]	99.6	0.020	yes	yes
466	*Faecalibacterium prausnitzii*	97.6	0.021	yes	yes
78	*Prevotella copri*	99.6	0.022	yes	yes
348	*Intestinibacter bartlettii*	100.0	0.024	no	yes
352	*Terrisporobacter petrolearius*	99.6	0.032	no	yes
318	*Bacteroides xylanolyticus*	98.0	0.035	yes	yes
390	*Clostridium disporicum*[3]	97.2	0.040	yes	yes
280	*Coprococcus eutactus*	99.6	0.042	yes	yes
399	*Catenibacterium mitsuokai*	97.6	0.044	yes	yes
519	*Campylobacter troglodytis*	98.4	0.047	no	no
Species with higher relative abundance in impaired language group					
219	*Granulicatella elegans*	100.0	0.0005	no	no
57	*Parabacteroides*	97.2	0.015	no	yes
31	*Bifidobacterium longum*[4]	99.6	0.027	yes	no
529	*Escherichia/Shigella*	99.6	0.028	yes	no
528	*Escherichia/Shigella*	100.0	0.034	yes	no
521	*Campylobacter coli*	97.2	0.041	no	no

A two-tailed test was performed for bacterial relative abundances of the gut microbiota of Ugandan children at the age of 24 months for language impaired (n = 61; BSID-III scores < 100) and non-impaired groups (n = 78; BSID-III scores ≥ 100).

Bacterial species [based on BLAST searches of amplicon sequence variants (ASV's)] listed in the table had a P-value below 0.05. Species with identity scores below 97% were excluded from the list in the table. Unambiguously assigned bacterial species are indicated by superscripts.

The ASV-match of Bifidobacterium catenulatum[1] is identical to that of Bifidobacterium pseudocatenulatum, Bifidobacterium kashiwanohense, Bifidobacterium tsurumiense, Bifidobacterium callitrichidarum, and Bifidobacterium gallicum (assigned to the catenalatum group); Bifidobacterium adolescentis[2] is identical to that of Bifidobacterium faecale and Bifidobacterium stercoris (assigned to the adolescentis group); Clostridium disporicum[3] identical to C. saudiense; Bifidobacterium longum[4] to B. breve (assigned to the longum group).

language development. Relative abundance of members of the *longum* group show at 24 months a negative correlation with language development, but a positive correlation at 36 months (**Figure 4**). At this point it is not clear why the relative abundance of the *B. longum* species group at 24 months, known to be beneficial and dominant in infants, correlate negatively with language development, in contrast to species from *B. catenulatum* and *adolescentis* groups, which are generally more prevalent in adults (Arboleya et al., 2016). Many of the bacterial species listed in **Table 2**, which are more abundant in the above average group, are known butyrate producers, including *C. eutactus*, *F. prausnitzii*, *Holdemanella biformis*, *Roseburia hominis*, *Clostridium disporicum*, and *Catenibacterium mutsuokai* (Vital et al., 2014). The SCFA butyrate has been implicated to play a role in brain function, as further discussed below.

Increased Predominance of Oxygen Tolerant Species in Children Impaired in Language Development

While 17 out of 19 ASV's with significant scores in the above average group matched with strictly anaerobic species (all except for *Campylobacter troglodytis*), we identified only one ASV matching with a strictly anaerobic bacterium (*Parabacteroides*) among the significant scores in the below average group (**Table 2**). This result is in line with the notion that a relatively high redox potential in the environment of the gut is an adverse condition for language development. It should be noted that majority of the ASV's in the below average group (4 out of 6) match to species with known adverse effects in humans, including *Granulicatella elegans* (**Table 2**). Although known to be part of the normal intestinal human microbiota, this species has often been implicated in adverse conditions. In addition, aerobic *Escherichia/Shigella*, and *Campylobacter coli* species are known as major foodborne pathogens, causing the widely occurring diseases shigellosis and campylobacteriosis, which lead to severe diarrhea, in particular at relatively high prevalence among children in the developing world. A box plot of the MAPI indices among all the children in both language development groups indicated a slight difference ($P = 0.09$) between the two groups (**Figure 5**).

DISCUSSION

The Value of Alternative Prediction Models

For the analysis presented in this paper, we identified promising predictors of language development in a field study from a large set of potential predictors that are likely to be correlated. Because of this correlation, it was imperative to use methods that could reveal alternative explanations of the same data. In the field data we studied, we indeed found substantial correlations among the observed microbiota abundances that could serve as potential predictors (4164 pairwise correlations were larger than 0.5).

Revealing alternative explanations of the same data requires the fitting of multiple models. We used MIO in our model search. The strong point of this approach is that one can impose constraints relevant for the data at hand. We used this option in our ranking of the second best down to 20th best models. Of particular use were the models with 3 and 4 predictors. There was a clear best 3-predictor model among 20 alternative models. This model included the language ability of the children at 24 months, the abundance of *C. eutactus* in microbiota taken at 24 months, and the abundance of *B. longum* in microbiota taken at 36 months. The fact that this model is clearly better than the alternatives suggests that we should include the three predictors mentioned in any case. However, there might still be additional predictors that could improve the model fit. This was investigated by fitting 4-parameter models as well.

There was no clear best 4-parameter model. However, *C. eutactus* abundance at 24 months was consistently present in all 4-parameter models, while the other two predictors in the best 3-parameter model were included in 14 of the 20 best 4-parameter models. By focusing on the common predictors present in the best models, we believe that we avoided overfitting the data. The remaining predictors were present in at most 5 out of the 20 best 4-predictor models. We conclude that there is no clear evidence favoring inclusion of a fourth predictor.

A further use of constraints in the MIO approach can help finding good models that include synergistic or antagonistic effects of the microbiota species. However, MIO is still limited in the size of the models it can handle. In particular, it is computationally infeasible to arrive at the best 5-term model based on 1170 potential model terms. As there are 1163 individual predictors involving microbiota composition, synergistic or antagonistic effects among the species would increase this number with $0.5 \times 1163 \times (1163–1) = 675{,}703$ further terms. It is infeasible to have a successful model search among this number of terms.

Importance of Early Life Acquirement of the Butyrate-Producing *Coprococcus eutactus* for Language Development

One of the most intriguing findings of this work is the correlation between the abundance of members of saccharolytic clostridia in the gut of Uganda children at 24 months with the composite score for language development of the children at 36 months. We identified *C. eutactus* (42 out of 60 models) and *I. bartlettii* (8 out of 60 models). They belong to the Lachnospiraceae and Peptostreptococcaceae, respectively, both families within the Clostridia, a class of obligatory anaerobic spore-forming bacteria. Both species produce SCFA's, the primary end-products of fermentation of non-digestible carbohydrates that become available to the gut microbiota and gut epithelial cells. The SCFA's are mainly produced through saccharolytic fermentation of carbohydrates. While *C. eutactus* is known to produce the SCFA's formate, acetate and butyrate (Holdeman and Moore, 1974), *I. bartlettii* produces the SCFA's isobutyrate and isovalerate (Song et al., 2004). It is well established that SCFA's, in particular butyrate, are important substrates for maintaining

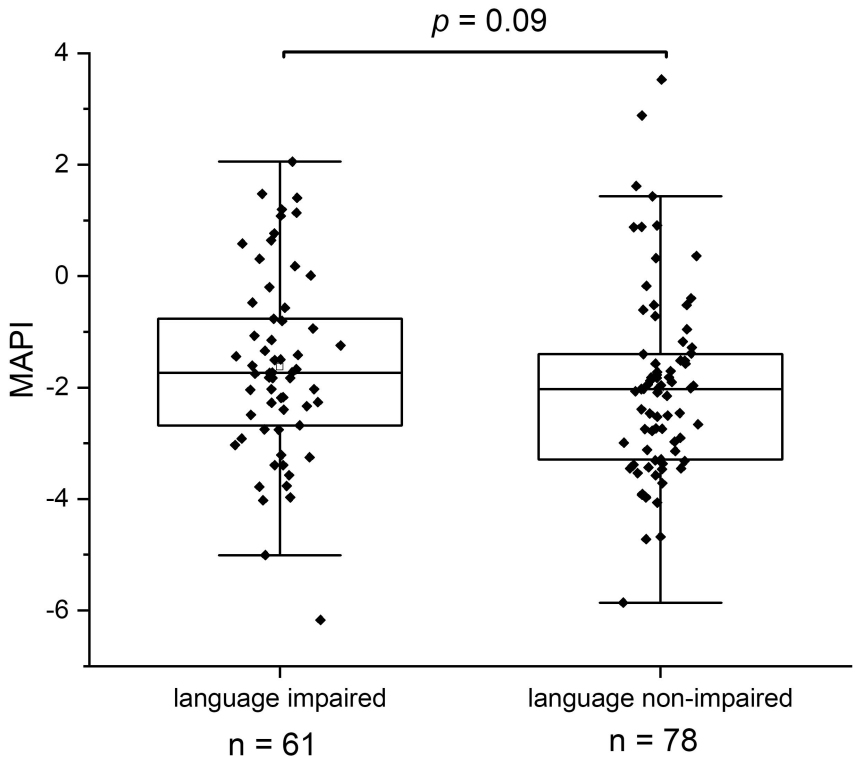


FIGURE 5 | The Metagenomic Aerotolerant Predominance Index (MAPI). The index is presented in box plots for the groups of language impaired (n = 61) and language non-impaired children (n = 78).

the colonic epithelium, elicit effects on lipid metabolism and adipose tissue at several levels, in appetite regulation and energy intake, and play a role in regulation of the immune system (Morrison and Preston, 2016). In addition, butyrate has been shown to protect the brain and enhance plasticity in animal models for neurological disease. In agreement with a role for the production of butyrate in the gut for improved language development, studies with animal models show that butyrate is able to reverse stress-induced decrease of neurotrophic factors and cognition impairment both at early and later stages of life (Valvassori et al., 2014). A number of mechanisms have been attributed to the beneficial role of butyrate in brain function, including its action as a histone deacetylase inhibitor and as an activator of G protein-coupled receptors (GPR's); a lower level of histone acetylation is a characteristic of many neurodegenerative diseases, and butyrate has been shown to activate GPR109a, potentially leading to anti-inflammatory effects in the brain (Bourassa et al., 2016).

The most consistent predictor in our MIO models for language development at 36 months was the abundance of *C. eutactus* in gut microbiota when the children were 24 months of age. This is in agreement with the concept of a maturation program with distinct phases of microbiota compositions, where earlier phases can affect health outcomes later in life (Backhed et al., 2015; Stewart et al., 2018). The dynamics of the relative abundance of *C. eutactus* was highlighted in a study on the human infant gut microbiome in development and in progression toward type 1 diabetes (Kostic et al., 2015). This longitudinal study indicated a maximum of *C. eutactus* relative abundance in healthy infants at approximately 24 months, while the abundance of *C. eutactus* type 1 diabetes predisposed children remained at constant, at relatively low levels in the first years of life. So far we only have analyzed the gut microbiota in children at 24 and 36 months in our cohort, thus at this moment we cannot yet make any substantiated statements about the longitudinal development of the gut microbiota in our cohort. However, the results in our study are in agreement with a model that holds that relatively high levels of *C. eutactus* at 24 months are beneficial, as they are present in the group of children with above average language development at 36 months.

A number of other uncertainties and limitations should be considered in the interpretation of our results. Among all hypervariable regions of 16S rRNA gene, the V4 region used in this study ranks first in sensitivity as a marker for bacterial and phylogenetic analysis (Yang et al., 2016). Nevertheless, these amplicon sequence libraries allow in some cases only a classification of microbiota members on the genus level. Therefore, we carefully examined all assignments to the species level in this study. Overall, the correlation between genomes of closely related species suggests that it may be effective to predict functions encoded in an organism's genome. A recent study showed phylogeny and function to be sufficiently linked that prediction of function from 16S rRNA gene amplicons can provide useful insights (Langille et al., 2013). However, in our

view metagenome sequencing to reveal the full genetic capacity of the gut microbiota, intervention studies with *C. eutactus* in a germ-free mouse model and *in vivo* metabolite measurements are required to acquire additional evidence on a beneficial role of butyrate production and additional neuroactive potential of the gut bacterium *C. eutactus* in cognitive development.

Relative Abundance of *Coprococcus eutactus* Correlates to Multiple Cognitive Outcomes

Interestingly, a recent study on the neuroactive potential of the gut microbiota with a large cohort (Flemish Gut Flora Project; n = ?1,054) revealed that butyrate-producing *Coprococcus* bacteria were consistently associated with higher quality of life indicators and depleted in depression (Valles-Colomer et al., 2019). The authors of this study performed a module-based examination of metabolic pathways by members of the gut microbiota in order to investigate its neuroactive potential. They observed that a gene encoding for the synthesis of 3,4-dihydroxyphenylacetic acid (a metabolite of the neurotransmitter dopamine) was strongly associated with the presence of *C. eutactus* and quality of life indicators. Notably, a second metabolic module, which co-varied with quality of life indicators in their cohort, is the synthesis of isovalerate. This ability to synthesize this SCFA happens to be present in *Intestinibacterium bartlettii* (Song et al., 2004), which is the species matching to ASV ID 348 in our best fitting models.

A further evaluation of the current scientific literature confirms that the relative abundance of the genus *Coprococcus*, and in particular the species *C. eutactus*, correlates with other cognitive outcomes. A lower relative abundance of *Coprococcus* was found in autistic patients compared to neurotypical controls (**Table 3**). An independent study confirmed lower levels of fecal acetic acid and butyrate in autistic subjects (Liu et al., 2019). A decreased relative abundance of *C. eutactus* was also observed in fecal samples and mucosal biopts from Russian and American patients with Parkinson's disease (PD), respectively (**Table 3**). In both studies, potentially anti-inflammatory, butyrate-producing genera, *Coprococcus*, *Faecalibacterium* and *Blautia* were significantly more abundant in feces of controls than PD patients, feeding the hypothesis that an altered gut microbiota could contribute to inflammation-induced development of PD pathology (Keshavarzian et al., 2015). A cross-sectional study on schizophrenia patients also indicated that the level of butyrate producing bacterial genera, including *Coprococcus*, *Blautia* and *Roseburia* significantly decreased in comparison to healthy controls. The observed differences in microbiota compositions were proposed as a basis for the development of microbiota-based diagnosis for schizophrenia (Shen et al., 2018). However, it is clear that among these differences, i.e., a decrease of a number

TABLE 3 | Correlations between *Coprococcus eutactus* and human mental health outcomes.

Genus species	Finding	Cohort	Sample size (n)	P-value	Statistical test	Study references
Coprococcus	Depleted in cohort participants with depression	Flemish Gut Flora project	1054	<0.05	Covariance test	Valles-Colomer et al., 2019
Coprococcus	Depleted in cohort participants with depression	Dutch lifeline DEEP	1063	<0.05	Covariance test	Valles-Colomer et al., 2019
Coprococcus	Lower relative abundance in autistic patients compared to neurotypical controls	American children (20 neurotypical and 20 autistic)	40	0.001	Mann–Whitney *U* test	Kang et al., 2013
Coprococcus	Lower relative abundance in Parkinson's-diseased patients compared to healthy controls	American adults (34 Parkinson's patients and 31 healthy controls)	65	0.03	Kruskal–Wallis test	Keshavarzian et al., 2015
Coprococcus eutactus	Lower relative abundance in Parkinson-diseased patients compared to healthy controls	Siberian adults (89 Parkinson's patients and 66 healthy controls)	157	0.03	White's *t*-test	Petrov et al., 2017
Coprococcus	Relative abundance reduced in schizophrenia patients	64 schizophrenia patients and 53 healthy controls	117	0.004	Principal coordinate analysis Welch's *t-test*	Shen et al., 2018
Coprococcus eutactus	Predictor in gut microbiota at 24 months for language development at 36 months	Rural Ugandan children	139	<0.001	All subsets regression	This study

The table includes information about cohort, sample size, statistical test, P-value, and study reference.

of butyrate-producing bacterial genera, a similar correlation can be observed for very different adverse cognitive outcomes, including the impaired language development with Ugandan children in our study.

We looked in this study for other overall differences between bacterial gut communities in the language impaired and language non-impaired groups of children and found higher levels of oxygen tolerant species in the first group. This finding concerns specific, potentially pathogenic species with significant higher relative abundance in the language impaired group (*G. elegans*, *Escherichia/Shigella*, *C. coli*), but also to slight differences in the overall MAPI index. As this index indicates an aerotolerant predominance for MAPI > 0 and anaerobic predominance for MAPI < 0, it is clear that both groups have an anaerobic predominance of bacterial species in the gut. Apparently, the increase of a number of oxygen tolerant species in the language impaired group is not so much reflected by the overall MAPI index. Possibly, this results from the fact that the Ugandan children in our study group are not severely malnourished, as they are on average moderately stunted ($-3 <$ HAZ < -2). More severe malnourishment could have led to the overall depletion of anaerobic bacteria and proliferation of oxygen tolerant bacteria, as shown in the gut microbiota of severely malnourished children (Million et al., 2016). In order to confirm the findings in this study, we propose to repeat the analysis and investigate cognitive development as a function of the MAPI index in a similar cohort accompanied by metabolite measurements in stool samples. In parallel, we propose to set up an intervention study aiming at the reduction of the gut redox potential as a stimulus to create a better growth environment for beneficial, strictly anaerobic gut bacteria, including *C. eutactus* and other butyrate producers identified in this study.

DATA AVAILABILITY STATEMENT

All 16S rRNA amplicon paired end reads of the gut microbiota samples sequenced in this study are accessible at BioProject PRJNA517509. All data presented in this study can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The study was approved by The AIDS Support Organization Research Ethics Committee (no. TASOREC/06/15-UG-REC-009) and by the Uganda National Council for Science and Technology (no. UNCST HS 1809) as well as by the Norwegian Regional Committee for Medical and Health Research Ethics (no. 2013/1833). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

ES developed the models, interpreted MIO outcomes, and made **Table 1** and **Figure 2**. RK and ES drafted the manuscript. AV carried out the MIO analysis and made **Figure 1**. RK made **Figures 3–5**. GM and PA carried out the field work in Uganda, collected and analyzed the data of the education intervention study. AWa supported in microbiota data analysis. AWe and PI designed the education intervention study and analyzed developmental data. JS carried out the PCR, Mann–Whitney *U* test, analysis for the core microbiota, and made **Tables 2**, **3**. WS supported in the translation of the results. All authors reviewed and approved the final manuscript.

FUNDING

This research was funded by the Vrije Universiteit Amsterdam (Amsterdam, Netherlands), and the Throne Holst Foundation (Oslo, Norway).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.681485/full#supplementary-material

REFERENCES

Albers, C., and Grieve, A. (2007). Test review: Bayley, N. (2006). Bayley scales of infant and toddler development– third edition. San Antonio, TX: Harcourt Assessment. *J. Psychoeduc. Assess.* 25, 180–190. doi: 10.1177/0734282906297199

Arboleya, S., Watkins, C., Stanton, C., and Ross, R. P. (2016). Gut bifidobacteria populations in human health and aging. *Front. Microbiol.* 7:1204. doi: 10.3389/fmicb.2016.01204

Atukunda, P., Muhoozi, G. K. M., Van Den Broek, T. J., Kort, R., Diep, L. M., Kaaya, A. N., et al. (2019). Child development, growth and microbiota: follow-up of a randomized education trial in Uganda. *J. Glob. Health* 9:010431.

Backhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 17, 690–703.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Ann. Stat.* 44, 813–852.

Bienenstock, J., Kunze, W., and Forsythe, P. (2015). Microbiota and the gut-brain axis. *Nutr. Rev.* 73(Suppl. 1), 28–31.

Bourassa, M. W., Alim, I., Bultman, S. J., and Ratan, R. R. (2016). Butyrate, neuroepigenetics and the gut microbiome: can a high fiber diet improve brain health? *Neurosci. Lett.* 625, 56–63. doi: 10.1016/j.neulet.2016.02.009

Cryan, J. F., O'riordan, K. J., Cowan, C. S. M., Sandhu, K. V., Bastiaanssen, T. F. S., Boehme, M., et al. (2019). The microbiota-gut-brain axis. *Physiol. Rev.* 99, 1877–2013.

de Boer, P., Caspers, M., Sanders, J. W., Kemperman, R., Wijman, J., Lommerse, G., et al. (2015). Amplicon sequencing for the quantification of spoilage microbiota in complex foods including bacterial spores. *Microbiome* 3:30.

Furnival, G., and Wilson, R. (1974). Regression by leaps and bounds. *Technometrics* 16, 499–511.

Gorvitovskaia, A., Holmes, S. P., and Huse, S. M. (2016). Interpreting *Prevotella* and *Bacteroides* as biomarkers of diet and lifestyle. *Microbiome* 4:15.

Gurobi, (2017). *Gurobi Optimization: Gurobi 8 Performance Benchmarks [Online]*. Available online at: http://www.gurobi.com/pdfs/benchmarks.pdf (accessed February 15, 2021)

Holdeman, L., and Moore, W. (1974). New genus, *Coprococcus*, Twelve new species, and emended descriptions of four previously described species of bacteria from human feces. *Int. J. Syst. Bacteriol.* 24, 260–277. doi: 10.1099/00207713-24-2-260

Jameson, K. G., Olson, C. A., Kazmi, S. A., and Hsiao, E. Y. (2020). Toward understanding microbiome-neuronal signaling. *Mol. Cell* 78, 577–583. doi: 10.1016/j.molcel.2020.03.006

Kang, D. W., Park, J. G., Ilhan, Z. E., Wallstrom, G., Labaer, J., Adams, J. B., et al. (2013). Reduced incidence of *Prevotella* and other fermenters in intestinal microflora of autistic children. *PLoS One* 8:e68322. doi: 10.1371/journal.pone.0068322

Keshavarzian, A., Green, S. J., Engen, P. A., Voigt, R. M., Naqib, A., Forsyth, C. B., et al. (2015). Colonic bacterial composition in Parkinson's disease. *Mov. Disord.* 30, 1351–1360. doi: 10.1002/mds.26307

Kort, R. (2019). *Fecal Microbiotia Compositions from Ugandan Children of 2 and 3 Years Old [Online]*. Available: https://www.ncbi.nlm.nih.gov/bioproject/517509 (accessed February 15, 2021)

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyotylainen, T., Hamalainen, A. M., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* 17, 260–273. doi: 10.1016/j.chom.2015.01.001

Langille, M. G., Zaneveld, J., Caporaso, J. G., Mcdonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Liu, S., Li, E., Sun, Z., Fu, D., Duan, G., Jiang, M., et al. (2019). Altered gut microbiota and short chain fatty acids in Chinese children with autism spectrum disorder. *Sci. Rep.* 9:287.

Mann, H., and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18, 50–60. doi: 10.1214/aoms/1177730491

Miller, A. (2002). *Subset Selection in Regression*. Boca Raton, FL: Chapman & Hall/CRC.

Million, M., and Raoult, D. (2018). Linking redox to human microbiome. *Hum. Microb. J.* 10, 27–32. doi: 10.1016/j.humic.2018.07.002

Million, M., Tidjani Alou, M., Khelaifia, S., Bachar, D., Lagier, J. C., Dione, N., et al. (2016). Increased gut redox and depletion of anaerobic and methanogenic prokaryotes in severe acute malnutrition. *Sci. Rep.* 6:26051.

Morrison, D. J., and Preston, T. (2016). Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* 7, 189–200. doi: 10.1080/19490976.2015.1134082

Muhoozi, G. K., Atukunda, P., Mwadime, R., Iversen, P. O., and Westerberg, A. C. (2016). Nutritional and developmental status among 6- to 8-month-old children in southwestern Uganda: a cross-sectional study. *Food Nutr. Res.* 60:30270. doi: 10.3402/fnr.v60.30270

Muhoozi, G. K. M., Atukunda, P., Diep, L. M., Mwadime, R., Kaaya, A. N., Skaare, A. B., et al. (2018). Nutrition, hygiene, and stimulation education to improve growth, cognitive, language, and motor development among infants in Uganda: a cluster-randomized trial. *Matern. Child Nutr.* 14:e12527.

Parker, M., Zobrist, S., Donahue, C., Edick, C., Mansen, K., Hassan Zade Nadjari, M., et al. (2018). Naturally fermented milk from northern senegal: bacterial community composition and probiotic enrichment with *Lactobacillus rhamnosus*. *Front Microbiol* 9:2218. doi: 10.3389/fmicb.2018.02218

Petrov, V. A., Saltykova, I. V., Zhukova, I. A., Alifirova, V. M., Zhukova, N. G., Dorofeeva, Y. B., et al. (2017). Analysis of gut microbiota in patients with Parkinson's disease. *Bull. Exp. Biol. Med.* 162, 734–737.

Rhee, S. H., Pothoulakis, C., and Mayer, E. A. (2009). Principles and clinical implications of the brain-gut-enteric microbiota axis. *Nat. Rev. Gastroenterol. Hepatol.* 6, 306–314. doi: 10.1038/nrgastro.2009.35

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34, 374–378. doi: 10.2144/03342mt01

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/aem.01541-09

Shen, Y., Xu, J., Li, Z., Huang, Y., Yuan, Y., Wang, J., et al. (2018). Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia: a cross-sectional study. *Schizophr. Res.* 197, 470–477. doi: 10.1016/j.schres.2018.01.002

Shetty, S. A., Hugenholtz, F., Lahti, L., Smidt, H., and De Vos, W. M. (2017). Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. *FEMS Microbiol. Rev.* 41, 182–199. doi: 10.1093/femsre/fuw045

Song, Y. L., Liu, C. X., Mcteague, M., Summanen, P., and Finegold, S. M. (2004). *Clostridium bartlettii* sp. nov., isolated from human faeces. *Anaerobe* 10, 179–184. doi: 10.1016/j.anaerobe.2004.04.004

Stewart, C. J., Ajami, N. J., O'brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* 562, 583–588.

Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E. F., Wang, J., Tito, R. Y., et al. (2019). The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* 4, 623–632. doi: 10.1038/s41564-018-0337-x

Valvassori, S. S., Varela, R. B., Arent, C. O., Dal-Pont, G. C., Bobsin, T. S., Budni, J., et al. (2014). Sodium butyrate functions as an antidepressant and improves cognition with enhanced neurotrophic expression in models of maternal deprivation and chronic mild stress. *Curr. Neurovasc. Res.* 11, 359–366. doi: 10.2174/1567202611666140829162158

Vazquez, A. R., Schoen, E. D., and Goos, P. (2020). A mixed integer optimization approach for model selection in screening experiments. *J. Qual. Technol.* doi: 10.1080/00224065.2020.1712275

Vital, M., Howe, A. C., and Tiedje, J. M. (2014). Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *mBio* 5:e00889.

Wacoo, A. P., Atukunda, P., Muhoozi, G., Braster, M., Wagner, M., Broek, T., et al. (2020). Aflatoxins: occurrence, exposure, and binding to *Lactobacillus* species from the gut microbiota of rural Ugandan children. *Microorganisms* 8:347. doi: 10.3390/microorganisms8030347

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/aem.00062-07

Wolters, M. A., and Bingham, D. (2011). Simulated annealing model search for subset selection in screening experiments. *Technometrics* 53, 225–237. doi: 10.1198/tech.2011.08157

Yang, B., Wang, Y., and Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* 17:135. doi: 10.1186/s12859-016-0992-y

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134. doi: 10.1186/1471-2105-13-134

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership