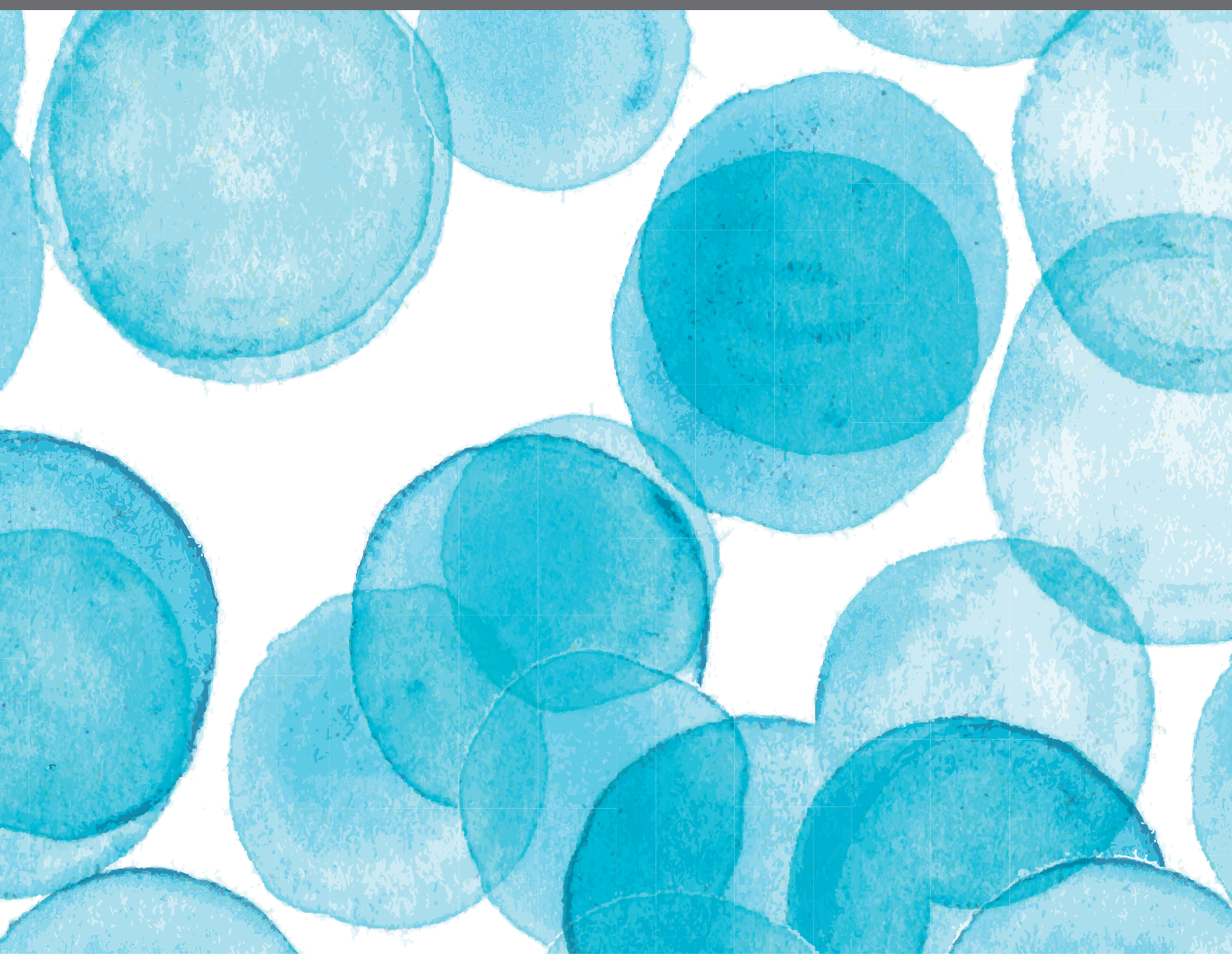
BIOINFORMATICS IN MICROBIOTA

EDITED BY: Xing Chen, Hongsheng Liu and Qi Zhao
PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88963-563-4
DOI 10.3389/978-2-88963-563-4

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: researchtopics@frontiersin.org

BIOINFORMATICS IN MICROBIOTA

Topic Editors:
Xing Chen, China University of Mining and Technology, China
Hongsheng Liu, Liaoning University, China
Qi Zhao, Shenyang Aerospace University, China

Citation: Chen, X., Liu, H., Zhao, Q., eds. (2020). Bioinformatics in Microbiota. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-563-4

Table of Contents

06 ***Editorial: Bioinformatics in Microbiota***
Xing Chen, Hongsheng Liu and Qi Zhao

09 ***The Genome-Scale Integrated Networks in Microorganisms***
Tong Hao, Dan Wu, Lingxuan Zhao, Qian Wang, Edwin Wang and Jinsheng Sun

20 ***PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine***
Balachandran Manavalan, Tae H. Shin and Gwang Lee

30 ***Identifying RNA N^6-Methyladenosine Sites in* Escherichia coli *Genome***
Jidong Zhang, Pengmian Feng, Hao Lin and Wei Chen

35 ***Faba Bean (*Vicia faba *L.) Nodulating Rhizobia in Panxi, China, are Diverse at Species, Plant Growth Promoting Ability, and Symbiosis Related Gene Levels***
Yuan X. Chen, Lan Zou, Petri Penttinen, Qiang Chen, Qi Q. Li, Chang Q. Wang and Kai W. Xu

45 ***Metformin Alters Gut Microbiota of Healthy Mice: Implication for its Potential Role in Gut Microbiota Homeostasis***
Wei Ma, Ji Chen, Yuhong Meng, Jichun Yang, Qinghua Cui and Yuan Zhou

53 ***Microbial Biogeography Along the Gastrointestinal Tract of a Red Panda***
Yan Zeng, Dong Zeng, Yi Zhou, Lili Niu, Jiabo Deng, Yang Li, Yang Pu, Yicen Lin, Shuai Xu, Qian Liu, Lvchen Xiong, Mengjia Zhou, Kangcheng Pan, Bo Jing and Xueqin Ni

63 ***Sc-ncDNAPred: A Sequence-Based Predictor for Identifying Non-coding DNA in* Saccharomyces cerevisiae**
Wenying He, Ying Ju, Xiangxiang Zeng, Xiangrong Liu and Quan Zou

72 ***PanGFR-HM: A Dynamic Web Resource for Pan-Genomic and Functional Profiling of Human Microbiome With Comparative Features***
Narendrakumar M. Chaudhari, Anupam Gautam, Vinod Kumar Gupta, Gagneet Kaur, Chitra Dutta and Sandip Paul

83 ***Human Microbe-Disease Association Prediction Based on Adaptive Boosting***
Li-Hong Peng, Jun Yin, Liqian Zhou, Ming-Xi Liu and Yan Zhao

92 ***Predicting Influenza Antigenicity by Matrix Completion With Antigen and Antiserum Similarity***
Peng Wang, Wen Zhu, Bo Liao, Lijun Cai, Lihong Peng and Jialiang Yang

101 ***Dietary Exposure to the Environmental Chemical, PFOS on the Diversity of Gut Microbiota, Associated With the Development of Metabolic Syndrome***
Keng Po Lai, Alice Hoi-Man Ng, Hin Ting Wan, Aman Yi-Man Wong, Cherry Chi-Tim Leung, Rong Li and Chris Kong-Chu Wong

112 ***PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method***
Yi Xiong, Qiankun Wang, Junchen Yang, Xiaolei Zhu and Dong-Qing Wei

121 *A Cross-Scale Neutral Theory Approach to the Influence of Obesity on Community Assembly of Human Gut Microbiome*
Wendy Li, Yali Yuan, Yao Xia, Yang Sun, Yinglei Miao and Sam Ma

129 *Human Microbe-Disease Association Prediction With Graph Regularized Non-Negative Matrix Factorization*
Bin-Sheng He, Li-Hong Peng and Zejun Li

140 *Evaluating the Value of Defensins for Diagnosing Secondary Bacterial Infections in Influenza-Infected Patients*
Siyu Zhou, Xianwen Ren, Jian Yang and Qi Jin

147 *Reconstruction and Analysis of a Genome-Scale Metabolic Model of* **Ganoderma lucidum** *for Improved Extracellular Polysaccharide Production*
Zhongbao Ma, Chao Ye, Weiwei Deng, Mengmeng Xu, Qiong Wang, Gaoqiang Liu, Feng Wang, Liming Liu, Zhenghong Xu, Guiyang Shi and Zhongyang Ding

159 *A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data*
Jian Xiao, Li Chen, Yue Yu, Xianyang Zhang and Jun Chen

171 *iVikodak—A Platform and Standard Workflow for Inferring, Analyzing, Comparing, and Visualizing the Functional Potential of Microbial Communities*
Sunil Nagpal, Mohammed Monzoorul Haque, Rashmi Singh and Sharmila S. Mande

186 *SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline*
Javier Tamames and Fernando Puente-Sánchez

196 *Exploring the Fecal Microbial Composition and Metagenomic Functional Capacities Associated With Feed Efficiency in Commercial DLY Pigs*
Jianping Quan, Gengyuan Cai, Ming Yang, Zhonghua Zeng, Rongrong Ding, Xingwang Wang, Zhanwei Zhuang, Shenping Zhou, Shaoyun Li, Huaqiang Yang, Zicong Li, Enqin Zheng, Wen Huang, Jie Yang and Zhenfang Wu

208 *Semen Microbiome Biogeography: An Analysis Based on a Chinese Population Study*
Zhanshan Ma and Lianwei Li

219 *Dynamic Development of Fecal Microbiome During the Progression of Diabetes Mellitus in Zucker Diabetic Fatty Rats*
Wen Zhou, Huiying Xu, Libin Zhan, Xiaoguang Lu and Lijing Zhang

236 *Web-gLV: A Web Based Platform for Lotka-Volterra Based Modeling and Simulation of Microbial Populations*
Bhusan K. Kuntal, Chetan Gadgil and Sharmila S. Mande

244 *Global Microbiome Diversity Scaling in Hot Springs With DAR (Diversity-Area Relationship) Profiles*
Lianwei Li and Zhanshan Ma

255 *Identification and Analysis of Human Microbe-Disease Associations by Matrix Decomposition and Label Propagation*
Jia Qu, Yan Zhao and Jun Yin

265 ***Artificial Neural Networks for Prediction of Tuberculosis Disease***
Muhammad Tahir Khan, Aman Chandra Kaushik, Linxiang Ji, Shaukat Iqbal Malik, Sajid Ali and Dong-Qing Wei

274 ***DMSC: A Dynamic Multi-Seeds Method for Clustering 16S rRNA Sequences Into OTUs***
Ze-Gang Wei and Shao-Wu Zhang

286 ***Identification of Phage Viral Proteins With Hybrid Sequence Features***
Xiaoqing Ru, Lihong Li and Chunyu Wang

298 ***A Novel Human Microbe-Disease Association Prediction Method Based on the Bidirectional Weighted Network***
Hao Li, Yuqi Wang, Jingwu Jiang, Haochen Zhao, Xiang Feng, Bihai Zhao and Lei Wang

311 ***Bacterial Community Succession, Transmigration, and Differential Gene Transcription in a Controlled Vertebrate Decomposition Model***
Zachary M. Burcham, Jennifer L. Pechal, Carl J. Schmidt, Jeffrey L. Bose, Jason W. Rosch, M. Eric Benbow and Heather R. Jordan

327 ***Application of Machine Learning in Microbiology***
Kaiyang Qu, Fei Guo, Xiangrong Liu, Yuan Lin and Quan Zou

337 ***Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model***
Dongmei Ai, Hongfei Pan, Xiaoxin Li, Yingxin Gao, Gang Liu and Li C. Xia

345 ***Evaluating the Effect of QIIME Balanced Default Parameters on Metataxonomic Analysis Workflows With a Mock Community***
Dimitrios Kioroglou, Albert Mas and Maria del Carmen Portillo

356 ***Metabolic Dependencies Underlie Interaction Patterns of Gut Microbiota During Enteropathogenesis***
Die Dai, Teng Wang, Sicheng Wu, Na L. Gao and Wei-Hua Chen

365 ***Assessing the Hybrid Effects of Neutral and Niche Processes on Gut Microbiome Influenced by HIV Infection***
Guanshu Yin and Yao Xia

372 ***RWHMDA: Random Walk on Hypergraph for Microbe-Disease Association Prediction***
Ya-Wei Niu, Cun-Quan Qu, Guang-Hui Wang and Gui-Ying Yan

382 ***Microbiota in Human Periodontal Abscess Revealed by 16S rDNA Sequencing***
Jiazhen Chen, Xingwen Wu, Danting Zhu, Meng Xu, Youcheng Yu, Liying Yu and Wenhong Zhang

394 ***Incorporating Statistical Test and Machine Intelligence Into Strain Typing of* Staphylococcus haemolyticus *Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry***
Chia-Ru Chung, Hsin-Yao Wang, Frank Lien, Yi-Ju Tseng, Chun-Hsien Chen, Tzong-Yi Lee, Tsui-Ping Liu, Jorng-Tzong Horng and Jang-Jih Lu

408 ***Cross-Regional View of Functional and Taxonomic Microbiota Composition in Obesity and Post-obesity Treatment Shows Country Specific Microbial Contribution***
Daniel A. Medina, Tianlu Li, Pamela Thomson, Alejandro Artacho, Vicente Pérez-Brocal and Andrés Moya

frontiers in Microbiology

EDITORIAL
published: 05 February 2020
doi: 10.3389/fmicb.2020.00100

Editorial: Bioinformatics in Microbiota

Xing Chen[1]*, *Hongsheng Liu*[2] and *Qi Zhao*[3]*

[1] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China, [2] School of Life Science, Liaoning University, Shenyang, China, [3] College of Computer Science, Shenyang Aerospace University, Shenyang, China

Keywords: bioinformatics, human microbiota, disease, big data, computational method

Editorial on the Research Topic

Bioinformatics in Microbiota

Microbiota are a group of microscopic organisms with simple structures and include bacteria, fungi, viruses, and others. Increasing numbers of biological experiments have shown that microbiota play a significant role in the occurrence and development of human diseases. Understanding the relationship between the microbiota and the host disease can be very useful in the treatment of complex diseases, such as inflammatory bowel disease, diabetes, and so on. However, using traditional wet experimental methods to identify microbe-disease associations is costly and time-consuming. During recent years, benefitting from the rapid development of artificial intelligence, machine learning, and new complex network techniques have been developed to work for the big data generated from human microbiome experiments. This Research Topic explores the potential for these computational methods applied to the research of human microbiota.

We are pleased to note that our Research Topic has attracted contributions from many highly regarded researchers in this field around the world, including from China, the USA, Spain, Chile, Korea, and India. We received 75 submissions, 39 of which were accepted for publication after rigorous reviews. We have further categorized these manuscripts into four subtopics with the Research Topic.

There are nine papers discussing the relationship between microbe and disease in the first part of this special issue. Zhou S. et al. examined the correlations between the gene expression levels of defensins and the viral and bacterial loads in the blood on a longitudinal, precision medicine study of a severe pneumonia patient infected by influenza A H7N9 virus. They showed that DEFB116 and DEFB127 are positively correlated and DEFB108B and DEFB114 are negatively correlated to the bacterial load. He B.-S. et al. proposed a novel predictive model of graph regularized non-negative matrix factorization for human microbe-disease relationship prediction based on known microbe-disease associations, Gaussian interaction profile kernel similarity for microbes and diseases, and symptom-based disease similarity. Wang et al. proposed a novel low-rank matrix completion model named MCAAS to infer antigenic distances between antigens and antisera based on partially revealed antigenic distances, virus similarity based on HA protein sequences, and vaccine similarity based on vaccine strains. Peng et al. established a model of adaptive boosting for human microbe-disease association prediction (ABHMDA) to reveal the associations between diseases and microbes. Chen J. et al. made a patient level analysis between abscess and healthy periodontium, which showed that *P. gingivalis*, and *Prevotella* spp. including *P. intermedia* were found to be dominant in the abscess of some patients compared to those of healthy periodontium, based on 16S rDNA metagenomic sequencing. Niu et al. introduced an *in silico* model named RWHMDA to predict underlying microbe-disease associations. Both cross-validation and case

OPEN ACCESS

Edited by:
Steve Lindemann,
Purdue University, United States

Reviewed by:
Wei Chen,
North China University of Science and Technology, China
Quan Zou,
University of Electronic Science and Technology of China, China

***Correspondence:**
Xing Chen
xingchen@amss.ac.cn
Qi Zhao
zhaoqi@lnu.edu.cn

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

Received: *25 October 2019*
Accepted: *17 January 2020*
Published: *05 February 2020*

Citation:
Chen X, Liu H and Zhao Q (2020) Editorial: Bioinformatics in Microbiota. Front. Microbiol. 11:100. doi: 10.3389/fmicb.2020.00100

studies on asthma, type 2 diabetes, and Crohn's disease revealed the reliability of RWHMDA. Li H. et al. proposed a novel prediction model called BWNMHMDA to accelerate the process of inferring potential microbe-disease associations, in which, the core idea is to construct a weighted bidirectional microbe-disease association network and then convert it into a matrix for correlation probability calculation. Qu J. et al. put forward the matrix decomposition and label propagation for human microbe-disease association prediction (MDLPHMDA) on the basis of the dataset of known microbe-disease associations collected from the database of HMDAD, the Gaussian interaction profile kernel similarity for diseases and microbes, and disease symptom similarity. Zhou W. et al. showed that the changes in the fecal microbiome were associated with age and disease progression in Zucker diabetic fatty rats.

Nine papers included in second part are focused on gut microbiota. Li W. et al. applied both Hubbell's and Sloan's neutral theory models to test the influence of obesity on the gut microbiome assembly from both community and species perspectives. Lai et al. investigated the effects of dietary perfluorooctane sulfonic acid (PFOS) exposure on gut microbiota in adult mice and examined the induced changes in animal metabolic functions. Zeng et al. characterized the microbial biogeographical characteristics in the GIT of a red panda using high-throughput sequencing technology. Ma W. et al. treated healthy mice with metformin and found that metformin could indeed prominently affect gut microbiome structure in healthy mice. Medina et al. compared the composition of the human gut microbiota of obese and lean people from six different regions and showed that the microbiota compositions in the context of obesity were specific to each studied geographic location. Yin and Xia firstly adopted a Silverman's test on the original results of the hybrid model, next using this strategy to reanalyze a dataset of HIV-related human gut microbiome in order to find HIV-specific changes in the assembly of gut microbial communities. Dai et al. constructed metabolic dependency networks using gut microbiota datasets of common enteric diseases including IBD and CRC, and revealed unappreciated interaction patterns of disease-enriched bacteria and probiotics. Ai et al. studied the microbial community structure of a CRC metagenomic dataset of 156 patients and healthy controls, and analyzed the diversity, differentially abundant bacteria, and co-occurrence networks. Quan et al. performed a comparative analysis of the fecal microbiota in DLY pigs with polarizing FE using 16S rRNA gene sequencing and shotgun metagenomic sequencing.

There are 12 papers with machine learning techniques applied in the research of microbiomes. Xiao et al. proposed a predictive framework to exploit sparse and clustered microbiome signals using a phylogeny-regularized sparse regression model. Xiong et al. developed a stacked ensemble model PredT4SE-Stack to predict T4SEs, which utilized an ensemble of base-classifiers implemented by various machine learning algorithms, to generate outputs for the meta-classifier in the classification system. Chaudhari et al. developed PanGFR-HM, a novel dynamic web-resource that integrates genomic and functional characteristics of 1,293 complete microbial genomes available from the Human Microbiome Project. He W. et al. collected the ncDNA benchmark dataset of *Saccharomyces cerevisiae* and developed a support vector machine-based predictor, called Sc-ncDNAPred, for predicting ncDNA sequences. Zhang et al. presented a computational method to identify m^6A sites in the *E.coli* genome by encoding the RNA sequences using nucleotide chemical properties and accumulated nucleotide frequency. Manavalan et al. described a novel computational method for predicting PVPs, called PVP-SVM, and utilized the available PVPs sequences to develop the method. Hao et al. reviewed three representative genome-scale cellular networks: GMN, TRN, and STN, and discussed the integration of the three types of networks. Qu K. et al. discussed the current application of machine learning methods in the microbiome. They reported that machine learning is widely used in microbiological research, and that it has focused on classification problems and analysis of interaction problems. Khan et al. developed an approach for prediction of the global burden of tuberculosis based on artificial neural networks. Ru et al. proposed a random forest method to classify bacteriophage virion proteins and non-phage virion proteins. Wei and Zhang presented a novel dynamic multi-seeds clustering method (namely DMSC) to pick operational taxonomic units. Chung et al. developed a statistical test-based method to determine the reference spectrum when dealing with alignment of mass spectra datasets, and constructed machine learning-based classifiers for categorizing different strains of *S. haemolyticus*.

Other studies are categorized as the fourth part of our special issue. There are nine papers in total in this part. Ma Z. et al. reconstructed a genome-scale metabolic model (GSMM) of a *Ganoderma lucidum* strain, and applied this model to elucidate detailed physiological characteristics and production of extracellular polysaccharide in this species. Chen Y. X. et al. isolated 65 rhizobial strains from faba bean, then studied their plant growth promoting ability with nitrogen free hydroponics, genetic diversity with clustered analysis of combined ARDRA and IGS-RFLP. Ma and Li analyzed the scaling of semen microbiome diversity across individuals with diversity-area relationship analysis, a recent extension to classic species-area relationship law in biogeography and ecology. Tamames et al. proposed a fully automatic pipeline (SqueezeMeta) for metagenomics/metatranscriptomics, covering all steps of the analysis. Nagpal et al. presented iVikodak, a multi-modular web-platform that hosts a logically inter-connected repertoire of functional inference and analysis tools, coupled with a comprehensive visualization interface. Kioroglou et al. performed metataxonomic analysis of two types of mock community standards with the same microbial composition for evaluating the effectives of QIIME balanced default parameters on a variety of aspects related to different laboratory and bioinformatic workflows. Burcham et al. monitored bacterial community structural and functional changes taking place during decomposition of the intestines, bone marrow, lungs, and heart in a highly controlled murine model. Li and Ma investigated the microbiome diversity scaling over space by analyzing the diversity-area relationship, which is an extension to classic species-area relationship law in biogeography. Kuntal et al. presented Web-gLV—a GUI based interactive platform for

generalized Lotka-Volterra (gLV) based modeling and simulation of microbial populations.

Finally, we want to thank all the authors who contributed their original work to our special issue and the reviewers for their valuable comments. We would like to express our sincere gratitude to the Specialty Chief Editor, Dr. Matthias Hess and Dr. George Tsiamis, and also the editorial office of Frontiers in Microbiology, for their excellent support and providing us with this opportunity to organize this hot topic issue successfully.

AUTHOR CONTRIBUTIONS

QZ and XC wrote and revised the manuscript. HL gave some helpful suggestions.

ACKNOWLEDGMENTS

This work was supported by Fundamental Research Funds for the Central Universities (2019ZDPY01).

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.

Copyright © 2020 Chen, Liu and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

REVIEW
published: 23 February 2018
doi: 10.3389/fmicb.2018.00296

The Genome-Scale Integrated Networks in Microorganisms

Tong Hao[1], Dan Wu[1], Lingxuan Zhao[1], Qian Wang[1], Edwin Wang[1,2] and Jinsheng Sun[1,3]**

[1] Tianjin Key Laboratory of Animal and Plant Resistance, College of Life Sciences, Tianjin Normal University, Tianjin, China, [2] Cumming School of Medicine, University of Calgary, Calgary, AB, Canada, [3] Tianjin Bohai Fisheries Research Institute, Tianjin, China

The genome-scale cellular network has become a necessary tool in the systematic analysis of microbes. In a cell, there are several layers (i.e., types) of the molecular networks, for example, genome-scale metabolic network (GMN), transcriptional regulatory network (TRN), and signal transduction network (STN). It has been realized that the limitation and inaccuracy of the prediction exist just using only a single-layer network. Therefore, the integrated network constructed based on the networks of the three types attracts more interests. The function of a biological process in living cells is usually performed by the interaction of biological components. Therefore, it is necessary to integrate and analyze all the related components at the systems level for the comprehensively and correctly realizing the physiological function in living organisms. In this review, we discussed three representative genome-scale cellular networks: GMN, TRN, and STN, representing different levels (i.e., metabolism, gene regulation, and cellular signaling) of a cell's activities. Furthermore, we discussed the integration of the networks of the three types. With more understanding on the complexity of microbial cells, the development of integrated network has become an inevitable trend in analyzing genome-scale cellular networks of microorganisms.

Keywords: integrated network, metabolic network, regulatory network, signal transduction network, microorganism

OPEN ACCESS

Edited by:
Xing Chen,
China University of Mining and Technology, China

Reviewed by:
Chunyu Zhu,
Liaoning University, China
Qinghua Cui,
Peking University, China

****Correspondence:***
Edwin Wang
edwin.wang@ucalgary.ca
Jinsheng Sun
jinshsun@163.com

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

Received: *23 December 2017*
Accepted: *08 February 2018*
Published: *23 February 2018*

Citation:
Hao T, Wu D, Zhao L, Wang Q, Wang E and Sun J (2018) The Genome-Scale Integrated Networks in Microorganisms. Front. Microbiol. 9:296. doi: 10.3389/fmicb.2018.00296

INTRODUCTION

With the development of bioinformatics and system biology, large-scale cellular network comes into the sight of researchers. Bioinformatics, based on data processing, model construction and theoretical analysis, integrates information from different molecular levels to understand how the biological system works. According to the types of biological information processing encoded in the network, the cellular networks have been classified into different types: genome-scale metabolic network (GMN), transcriptional regulatory network (TRN), and signal transduction network (STN). The most well-studied large-scale biological network is GMN, which is a fundamental framework in systems metabolic engineering (Kim et al., 2015). With the first GMN constructed for *Haemophilus influenzae* Rd (Edwards and Palsson, 1999), the current GMN allows systematic level predictions of metabolism in a variety of organisms (Yilmaz and Walhout, 2017). The main concept of transcriptional control was established in bacterial system by Jacob and Monod (1961). In the past decades, the development of genomic technology and computational biology promotes the construction of large-scale TRNs (Brent, 2016). The TRN is composed of the interactions between different transcriptional factors (TFs) and target genes. A TF, which

is encoded by a gene itself, may influence the expression of one or more target genes, which may subsequently give rise to the expression change of a serial of proteins or genes. The STN is different from the TRN in network structures and timescales. The STN contains protein–protein and protein–gene interactions, which includes multiple routes of rapid cell response to the external stimuli, whereas the TRN may need to produce sustained patterns of cellular activity over time (Babu et al., 2004; Papin et al., 2005). On the other hand, some proteins in the STN are TFs, which indicates some genes/proteins are in common between STNs and TRNs. The detailed comparisons of these networks have been described in the review (Wang et al., 2007).

From a system point of view, different kinds of biological networks are not working alone, but cooperate with each other to undertake their functions. Integrated network studies will build a more realistic model by investigating the interacting relationships and interacting effects among organism's different information processing components in its system. This kind of models has an important sense to the theoretical research of living systems and the construction of genetic engineering strains (Wang et al., 2010). In this article, we discussed the research progress about the integrated networks in microorganisms.

CELLULAR NETWORK

Cellular network analysis has become a hot research area in bioinformatics and system biology; it utilizes computer model and experimental data to analyze complex biological system in a global view, and offers guidance and expectation for *in vivo* experiments (Wu and Ma, 2014). Due to the complexity of the biological system, researchers have classified cellular networks into GMN, TRN, and STN based on the types of information processing of biological molecules.

Genome-Scale Metabolic Network

Due to the advances of genome sequencing, high-throughput data have been rapidly produced, which drives a transition from the traditional biology research. On the basis of genome sequencing and annotations in huge amounts of data, metabolic network reconstruction in a genome-scale has been developed rapidly (Francke et al., 2005; Notebaart et al., 2006). Currently, GMN has become an indispensable tool for studying the biological metabolic system (Pal et al., 2006; Feist and Palsson, 2008). It has important applications on designing classic paths of metabolic engineering, inverse metabolites synthesis, metabolic flux analysis, evolution analysis of metabolic pathways between different species, mining omic data, and identifying of the marks in enzyme engineering (Soh and Hatzimanikatis, 2010). GMN construction is based on genomic sequences, combining with genes, enzyme reactions, metabolic databases and related experimental data, to quantitatively study the metabolic processes of living organisms from a systematic perspective. All biochemical reactions in the cell have been included as a network and the GMN reflects the interactions between all the compounds involved in the metabolic processes and all the catalytic enzymes. The construction of a GMN allows an in-depth functional analysis of the biological metabolic system, which is different from the traditional approach analysis or biological response analysis, but try to understand the whole metabolic system from the systematic view. GMN brings a more comprehensive and accurate insight into cell metabolism of the whole system and the interaction relationships between different metabolic processes. On the other side, the topology of the metabolic networks among many organisms can reflect the dynamics of the metabolic system evolution, which can help us understand the history of life evolution in the context of metabolism (Ravasz et al., 2002; Stelling et al., 2002; Zhao, 2008; Deyasi et al., 2015). In all the genome-scale biological networks, GMN is the most extensive and deepest studied network, with its construction procedures generally normalized in Palsson's review (Thiele and Palsson, 2010). The process of constructing of a metabolic network mainly consists of four parts, including data collection, relationship model establishment, data curation, and transformation into a mathematical model (Thiele and Palsson, 2010). To date, the construction of metabolic network has been able to realize some degree of automation, and therefore, 100s of metabolic networks in different organisms have been constructed (Hao et al., 2012).

Genome-scale metabolic network can be used to simulate the growth of organisms. Among the GMNs, the most accurate, comprehensive and classical model in microorganisms is the GMN of *Escherichia coli* named *i*JO1366, which was constructed by Palsson's group in 2011. The model achieved 67.7 and 96% accuracies for the prediction of essential and non-essential genes in *E. coli*. It is capable of simulating the growth of *E. coli* on 334 kinds of nutrients (Orth et al., 2011). Recently, a novel updated GMN of *Clostridium difficile* which called *i*CDF834 has been presented. This network was constructed based on the model *i*MLTC806cdf and transcriptome data, which detailed the gene expression of the bacteria in various environments. It is worth mentioning that the synonymous codon usage bias was introduced into the model to remedy the inconsistence between gene expression and protein abundance, which is the first time that codon has been integrated into a GMN. The model achieved a quite high (92.3%) accuracy in predicting gene essentiality (Kashaf et al., 2017).

The GMN can be used to guide the metabolic engineering experiments. Using *Bacillus subtilis* as an example, Hao et al. (2013) constructed a GMN of *B. subtilis,* named *i*BSU1147. The model has been used to successfully predict the yields of four industrial products produced by *B. subtilis* [i.e., riboflavin, (R,R)-2,3-butanediol, cellulase Egl-237, and isobutanol]. The results have provided important guidance for the *in vivo* experiments (Hao et al., 2013). Recently, Piubeli et al. (2018) constructed a GMN *i*FP764 of halophilic bacterium *Chromohalobacter salexigens* to explore the cell factory for producing ectoine. This model was constructed based on the experimental data, genome sequences and re-annotation of metabolic genes. The GMN is capable of simulating the metabolic situation of *C. salexigens* in low and high yield of ectoines. The salinity-specific essential genes and the patterns of correlated reactions in central carbon and nitrogen metabolisms response to the change of salinity

were also simulated. The network is a useful tool to improve the production of ectoines with bacteria (Piubeli et al., 2018).

The GMN also has an important value for drug discovery. Chen et al. (2015) constructed a GMN of *Treponema pallidum*. *T. pallidum* has a very specific metabolic network compared to those of other bacterial pathogens. It lacks the oxidative phosphorylation tricarboxylic and acid cycle pathways as well as is incapable of synthesizing enzyme cofactors, fatty acids, and most amino acids. By analyzing topological structure and minimal cut sets of the network, they found that some hub reactions in pyrimidine and purine metabolisms play significant roles in *T. pallidum*, which may be helpful drug targets in the treatment of syphilis, a sexually transmitted infection caused by the *T. pallidum* (Chen et al., 2015). In the same year, Steinway et al. (2015) constructed a GMN of intestinal bacteria based on experimental data. This network summarized the relationships between clindamycin and clostridium infection. Based on the analysis of topological and chemical properties of the network, the drug targets could be screened using the GMN, which can be used in the design of the drug-molecule model (Cong, 2010) and subsequently be applied in the treatment of anticlostridium. They verified that *B. intestinihominis* can indeed slow the growth of *C. difficile* through *in vitro* experimental validation (Steinway et al., 2015).

Theoretically speaking, the number of completed genome sequenced species should be as same as the number of corresponding GMNs. However, the current number of GMNs is much less than the number of sequenced species. The main reason is that the network construction pipeline still needs manual proofreading procedures due to the imperfect genetic annotation algorithm. In addition, the incomplete understanding of biochemical mechanisms also affects the development of metabolic networks (Wang et al., 2010).

Genome-Scale Gene Transcriptional Regulatory Network

Gene transcriptional regulation is the most basic and important regulation mechanism in organisms. Therefore, computational analysis of the gene transcriptional regulation is helpful for the understanding of the interactions between transcriptional processes and TRNs, and could provide support for the understanding of the mechanisms of biological activities (De-nan, 2014).

The basic components of TRNs are the interactions between transcription factors (TFs) and the related target promoters which function in the activation or repression of gene transcription. In this definition, the intracellular signals that regulate TF activities or any other additional mechanisms that may influence the expression of genes were excluded, as well as the upstream environmental. Although the development of TRN is not as mature as that of GMN, the current TRN construction is more and more standardized and automated. The detailed construction method of the TRN in microorganism can be seen in this paper (Feist et al., 2009). The network construction method is roughly divided into four steps: Step 1: an automated genome-based construction with automated procedures and applying automated tools, such as SMILEY algorithm, GapFind/GapFill, and PathoLogic; Step 2: construction of the TRN based on bibliomic data or high-throughput data; Step 3: transforming a genome-scale reconstruction of the interactions into a computational model; Step 4: curation the network by adding physiological or *in vivo* experimental information to the genes and the network.

Transcriptional regulatory network is a very complex non-linear system. Therefore, it is difficult to be described in a mathematical model. So far, the studies of the TRN are still in the exploration stage in many aspects, and scientists are constantly exploring new and better ways to construct a more complete TRN. Using *Bacillus* as an example, in Sierro et al. (2008) improved the database of transcriptional regulation in *B. subtilis* (DBTBS), which is constructed in 1999 for collecting the information of experimentally characterized TFs, and they nearly doubled the information in DBTBS. Freyre-Gonzalez et al. (2013) examined each regulatory element that constituted the TRN of *B. subtilis* and presented some lessons from the construction processes. Arrieta-Ortiz et al. (2015) used the TRN of *B. subtilis* to calculate the activity of TFs with a new combination of composition analysis based on a large number of known transcriptome data and experimental data of *B. subtilis*. They predicted 2258 new regulatory interactions and recalled 74% previously known interactions with this model. The accuracy of predicted new regulation edges was 62% (391/635) (Arrieta-Ortiz et al., 2015). Faria et al. (2016) expanded a TRN for the central metabolism of *B. subtilis* reconstructed in 2008 by integrating the regulation information in DBTBS. They demonstrated that atomic regulons (ARs), which are the sets of genes with the same expression profile, are the effective references for improving the regulatory networks by finding the closely correlated genes in the ARs. The expanded model contains the regulatory information for 2500 of the 4200 genes in *B. subtilis* 168 (Faria et al., 2016). In addition, Gui et al. (2012) searched for the homologous TFs and their regulatory genes in the genetically closest pattern bacteria – *B. subtilis*, and used comparative proteomics to forecast a regulatory networks of *Bacillus pumilus*, which contains 195 TFs and 1201 controlled genes. The results of their study showed that comparative genomics is a reliable method to speculate the gene regulation network of some species based on the gene transcriptional regulatory relationships of their genetically close organism, which is the best and a widely studied model organism. This method offers a feasible way to explore some organisms' regulatory networks without large-scale gene expression data (Gui et al., 2012).

The TRN can also be used to treat the human disease. Recently, Fowler and Galan (2018) built a regulatory network of *Salmonella typhi*, a pathogen causing typhoid fever. Typhoid fever, which is a frequently happened disease in human, was mainly caused by the typhus toxin secreted by *S. typhi*. Typhoid fever toxin is expressed uniquely by intracellular bacteria with unknown regulatory network. Fowler and Galan (2018) built the TRN of *S. typhi* and developed an algorithm called FAST-INSeq to identify the genes and mutants which influence the expression of typhoid toxins.

This network can help to understand the expressional regulation of typhoid fever toxin in *S. typhi*, which would contribute to the treatment of typhoid fever (Fowler and Galan, 2018).

Genome-Scale Signal Transduction Network

Signal transduction is an important cellular activity, a living cell can recognize, connect and interact with each other through signal transduction pathway, and realize the overall functional coordination and unity. Signal transduction carries plenty of biological functions, and is closely connected with the development of many diseases (Liu et al., 2008). In the early years, scientists believed that the STN is a linear cascade of information transmission and amplification. However, due to more studies of the system, scientists found that the concept mentioned above is incorrect. Therefore, a new view taking a STN as a system consisting of multiple complicated elements interacting in a multifarious fashion emerges. This view conflicts with the protein-centric or single-gene approach commonly used in the traditional research (Levchenko, 2003). Scientists found that except a few STNs that contain fewer signals and simpler network structures, such as Jak-STAT pathway, most STNs are fairly complex (Papin and Palsson, 2004). In the cellular signaling system, a large amount of phosphorylation and dephosphorylation reactions makes the signal transduction process usually reversible. The lacking of mass flow and the complexity of network state changes make the STN different from the GMN and TRN.

To determine the relationships between the mechanism and molecular regulations in STNs, it requires a large number of experiments. However, the standard single cell technique contributes little to the STN because the states of signal change dynamically and are different between individual cells (Kamps and Dehmelt, 2017). Fortunately, computational approaches such as bioinformatics analysis using known data and biological knowledge can help to interpret the STN (Shlomi et al., 2006). As early as Gomez et al. (2001) used a statistic model to calculate the molecular interactions in *Saccharomyces cerevisiae* on the basis of protein structure domain and network topology. This method can generate potential signaling pathways and also be applied to multiple species (Gomez et al., 2001). Rother et al. (2013) summarized the approaches of constructing a STN and classified them into three types: network topology-based method where network simulation could be applied using Boolean models, network specific-state based method where the network is simulated using differential equation models, and reaction-contingency based method where the network is simulated using agent based models, site-specific logical models or bipartite Boolean models (Rother et al., 2013). Each of the three methods performs well in small network modules. However, when the scale of network extended to the genome level, none of them is perfect for dealing with the whole information in the entire STN (Le Novere et al., 2009). In recent years, lots of small-scale STNs has been studied, such as the STN of HRas (Herrero et al., 2017), mTOC1 (Hoxhaj et al., 2017), cell circle (Wang et al., 2018), and cellular adhesion (Zheng et al., 2014). At the meantime, much more efforts are being made to construct large-scale STNs. Therefore, it is challenging to model the large STNs. Even though signaling network in bacteria is not as complex as those in eukaryotes, the construction of a large-scale STN is still a major challenge. Vinayagam et al. (2011) constructed a protein–protein interaction network to resembling the signal transduction flow between 1126 proteins, in which the interactions were obtained from yeast two-hybrid experiments of more than 450 signaling proteins. This network has been used to predict 18 previously unknown modulators in EGF/ERK signaling. Their results shows that the integration of genetic experiments and the computational approach is valuable for elucidating interactions between signaling proteins and facilities the identification of proteins in STNs (Vinayagam et al., 2011). Wang et al. (2011) also performed an approach called CASCADE_SCAN to construct STN with high-throughput data, which further showed that the high-throughput experiments are becoming a powerful tool for assisting in reconstructing large-scale STNs. Besides, the integration of different techniques such as optogenetics, protein design, surface patterning, and chemical tools was reported to provide some valuable information of the dynamic state of signals in the network and contribute in the construction of large-scale STNs (Kamps and Dehmelt, 2017).

INTEGRATED NETWORKS IN MICROORGNISMS

The establishment of various biological networks simulates and validates key activities in cells. With the recent advances in high-throughput studies, it has been realized that it is necessary to integrate different levels of biological information processing networks to fully investigate the biological mechanisms of the organisms (Kitano, 2002; Ryll et al., 2014). Therefore, the integrated network based on different network types has become a trend in the field of system biology and bioinformatics.

Integrated Metabolic-Regulatory Networks

Metabolism and transcriptional regulation are two closely related cellular activities. Metabolites (substrates or reaction products) involved in metabolic reactions affect the activities of certain TFs or signal transduction pathways. On the other hand, enzyme-catalytic metabolic reactions are regulated by other genes or proteins, and the expression of enzymes is different in different environmental conditions. In recent decades, the integrative modeling of metabolic-regulatory networks has become an important research area in the modeling of microorganisms (Imam et al., 2015).

Covert et al. (2004) reconstructed the first genome-scale metabolic-regulatory integrated network of *E. coli* (*i*MC1010) based on the information derived from literature and databases. The network contains 906 metabolic genes and 104 regulatory genes, which regulate the expression of about 53% genes (479/906) in the *E. coli* metabolic network. This model is

capable of predicting the previously unknown TFs, which play important roles in regulating metabolic processes, and interactions between metabolites and TFs (Covert et al., 2004). In 2005, they further used the literature-curated network *i*MC1010v1 to evaluate the performance of the functional states calculated in 15,580 growth environments for *E coli*. The results showed that the TRN responds mainly to the electron acceptors, which agrees with known experimental data. They also found that a complicated network had a small amount of dominant modes and the network clusters of activity profiles can be organized based on the activities of a few TFs. The integrated network gives crisper references than the single metabolic network for the further experiments to determine the functional states of an organism (Barrett et al., 2005).

Goelzer et al. (2008) reported a manually curated metabolic-regulatory integrated network of *B. subtilis*. The network includes post-translational regulations translational regulation, and modulation of enzymatic activities in the central metabolism. They decomposed the complex network into different locally regulated modules and found that these modules were managed by global regulators. Their results exhibited the functional organization of the metabolic-regulatory integrated network of *B. subtilis* (Goelzer et al., 2008).

Chandrasekaran and Price (2010) proposed an algorithm named probabilistic regulation of metabolism (PROM) and constructed a genome-scale regulatory-metabolic integrated network model for *E. coli* and *Mycobacterium tuberculosis*. Before this effort, another method named regulatory flux balance analysis (rFBA) has been used to integrate transcriptional regulatory with metabolic networks. rFBA used the Boolean logic to link transcriptional control to the metabolic process, which permits only on/off states of the network components (Shlomi et al., 2007). PROM introduces probabilities instead of Boolean rules to represent gene expression and the interactions between gene and TF (Simeonidis et al., 2013). The analysis of integrated *E. coli* network demonstrates that metabolic-regulatory integrated network is more accurate and comprehensive than the models constructed based on manual curation of literature. The integrated *M. tuberculosis* model incorporated data from more than 2,000 TF, 1,300 microarrays, 1,905 KO phenotypes and 3,300 metabolic reactions. The application of PROM on this model shows the capability of PROM on various organisms. Particularly, they demonstrated the outstanding capability of PROM in predicting the cellular phenotypes, drug targets, and functions of less studied regulatory genes.

Jiang et al. (2012) constructed a metabolic-transcriptory integrated network of *Corynebacterium glutamicum* by combining public databases and literature databases. The network contains 1,384 reactions, 1276 metabolites, 88 regulators, and 999 transcriptional regulations. The study systematically reorganized and analyzed the transcriptional regulation information of *C. glutamicum*, and expanded it to the metabolic network. They also preliminarily analyzed the metabolic network of *C. glutamicum* on the basis of the bow-tie structure of the network (Ma and Zeng, 2003). This work showed that the integration of the TRN and the metabolic network with the gene-enzyme-reaction relationship could be the foundation for the large-scale data integration and simulation analysis. The advantages of this integrated network are the discoveries of the relationships between transcription and metabolism in cells, which can't be achieved if using either metabolic network or TRN only (Jiang et al., 2012).

Wang Z. et al. (2017) performed another algorithm called Integrated Deduced And Metabolism (IDREAM) to construct enhanced metabolic-regulatory integrated networks. IDREAM integrated Environment and Gene Regulatory Influence Network (EGRIN) models with the PROM framework. IDREAM performs better than PROM in the prediction of the phenotype and genetic interactions between TFs and metabolic processes in *S. cerevisiae* (Wang Z. et al., 2017).

Currently, large-scale metabolic-regulatory integrated network has been constructed for several microorganisms such as *E. coli* (Chandrasekaran and Price, 2010), *S. cerevisiae* (Herrgard et al., 2006), *Helicobacter pylori* (Schilling et al., 2002), *Phaeodactylum tricornutum* (Levering et al., 2017), comma shaped gram negative anaerobic bacteria (Mahadevan et al., 2006) and *C. glutamicum* (Kromer et al., 2004). Integration of metabolism and transcription processes is generally quite straightforward. Metabolic network produces precursors to synthesize the metabolites such as nucleotides and amino acids which are required by transcription processes. On the other hand, the TRN couples back to the metabolic network by managing the expression of the enzymes in the metabolic network and thus regulating the flux distribution among different metabolic functions (Feist et al., 2009).

Integrated Regulatory-Signaling Networks

The integration of microbial transcriptional regulatory and signaling network is still in the preliminary stage. Wang and Chen (2010) combined the transcriptional regulation and signal transduction pathway (e.g., mainly presented in the form of protein–protein interaction) to construct the integrated yeast cellular network. The network connects these two networks together to form an integrated network using the nodes (i.e., TFs) between the TRNs and signaling pathways. The integrated cellular networks related to heat shock, hyperosmotic stress, and oxidative stress were constructed and the connections between these networks were further analyzed. With the hyperosmotic stress related network, the highly connected hubs related to the stress response were predicted. The analyses of these networks have identified a few TFs to serve as the core in the bow-tie structure and the essential elements for the rapid response to stress. In addition, they also identified a couple of genes/proteins related to stress responses or potential drug targets. This method, however, only integrates the transcriptional regulatory data with the protein–protein interaction in the signal transduction pathways, but not the completed STN. In order to get a more complete integration, it also needs to list all the components in a STN, and then combined with the TRN for the integration (Wang and Chen, 2010). Recently, Ignatius Pang et al. (2018) construct another regulatory-signaling integrated network of

S. cerevisiae with protein–protein interaction as the bridge to link the regulatory (TF-gene pairs) and signaling (kinase-substrate pairs) parts. This network was used to investigate the negative genetic interactions and the genes in the negative genetic interactions closely related to the toxicity (Ignatius Pang et al., 2018).

In the study of algorithms, Roy et al. (2013) proposed a method called MERLIN (Modular regulatory network learning with per gene information) to reconstruct the regulatory network by identifying the connections from regulators, including proteins and TFs, to target genes. The regulatory network constructed by MERLIN actually reflects the integration of transcriptional regulation and signaling networks. The application of MERLIN on *S. cerevisiae* captured the co-regulatory relationships between downstream TFs and signaling proteins, and therefore uncovering the upstream signaling systems which control transcriptional responses (Roy et al., 2013). With the investigation of the integrated network, the regulation program of each gene in the human cells is much clearer than the application of either individual TRN or STN.

Integrated Metabolic-Signaling Networks

The development of integrated network for metabolic and signaling networks is still in the very beginning stage. Few metabolic-signaling integrated networks have been published. Imam et al. (2015) discussed the challenges in the integration of these two network types. Firstly, signaling mechanisms are closely related to the specific concentrations of related molecules, while constraint-based approaches widely used in metabolic network analysis cannot reflect the metabolite concentrations. Secondly, lots of kinetic parameters are required in the construction of dynamic quantitative signaling network, but these parameters are rarely available. This aspect limits the integration of metabolic and signal transduction. Boolean or stoichiometric methods which do not require kinetics parameters or metabolite concentrations might be a possible choice for the integration of metabolic and signaling networks in the future.

Integrated Metabolic-Regulatory-Signaling Networks

The integration of metabolic-regulatory-signaling networks is a challenge issue in the study of integrated networks. On the graphic view, there are common components (proteins or TFs) in metabolic, regulatory, and signaling networks (**Figure 1**). Therefore, it is theoretically possible to merge these three types of cellular networks into one integrated network. While actually, lots of elements should be considered in the integration process, such as the logics and computability. On a small scale network integration, Covert and Palsson (2002) developed a method named integrated FBA (iFBA) to model the dynamic behavior among metabolic, signaling, and regulatory networks. This method combines FBA with ordinary differential equations (ODE) and regulatory Boolean logic (**Figure 2**). They used this approach to construct an integrated network model of *E. coli* which combines a FBA based central carbon metabolic-regulatory network with an ODE based model of carbohydrate-uptaking-controlling network. They compared the prediction of *E. coli* single gene perturbation disturbance phenotypes and wild-type for diauxic growth on glucose/glucose-6-phosphate and glucose/lactose using rFBA and ODE methods. They found that iFBA is capable of identifying the dynamics of three transporters and three internal metabolites which cannot be predicted by

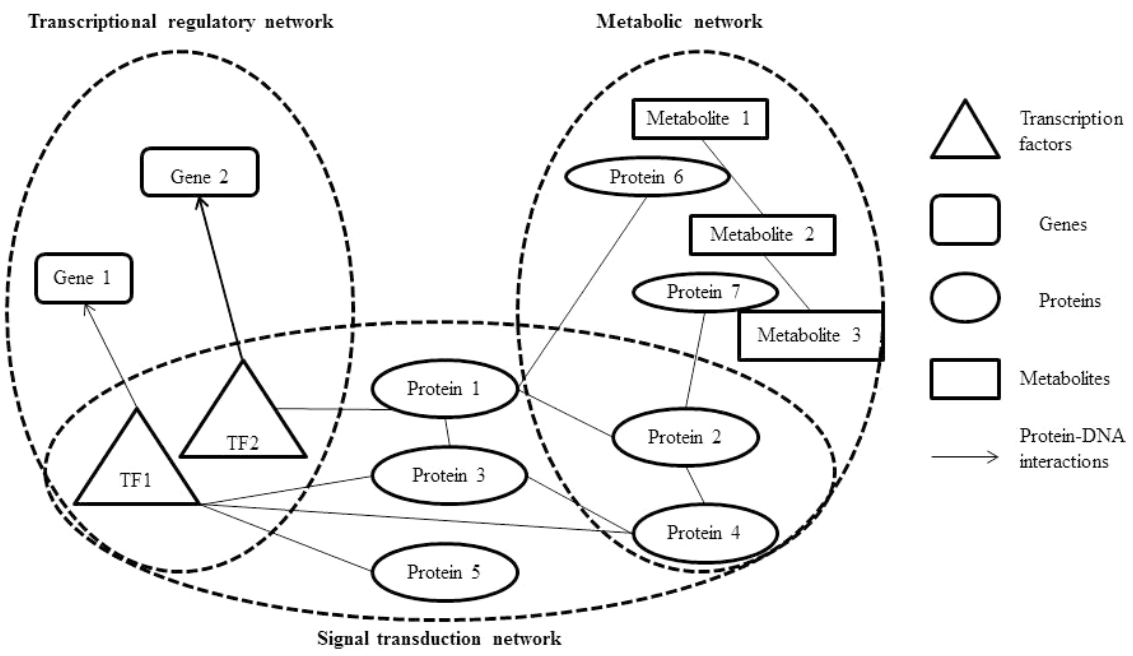


FIGURE 1 | Graphic view of the integrated cellular network.

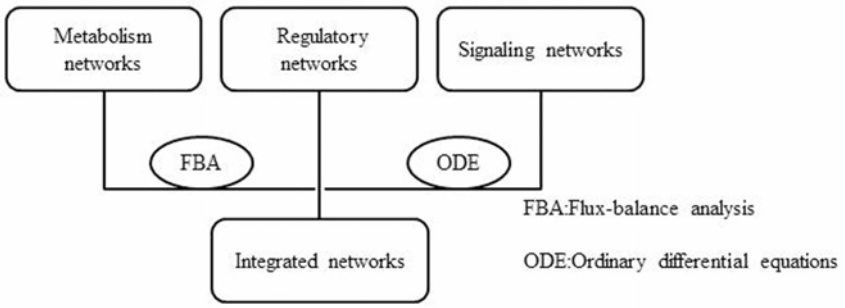


FIGURE 2 | Schematic diagram of iFBA.

rFBA alone. Furthermore, iFBA obtained different and more accurate phenotype predictions in the wild-type simulations and single gene perturbation simulations than the ODE model, which indicates that iFBA is an improvement over either individual rFBA or ODE method in network integration (Covert et al., 2008).

Lee et al. (2008) proposed a method called integrated dynamic FBA (idFBA) which could dynamically simulate cellular phenotypes with integrated networks. idFBA was applicable for the analysis of the integrated stoichiometric network of metabolic, regulatory, and signal transduction processes. In this method, the quasi-steady-state conditions were assumed for "fast" reactions and then the "slow" reactions was incorporated into the stoichiometric equation (**Figure 3**). idFBA has been applied to a representative small network of *S. cerevisiae*, in which metabolic, regulatory, and signaling activities have been included. Finally, idFBA got similar results with an equivalent kinetic model in the prediction of the influence of the extracellular environment on the cellular phenotypes. The advantage of idFBA is that it is capable of solving a linear programming problem without the detailed kinetic parameters, which makes it a possible approach for the genome-scale integration of metabolic, regulatory, and STNs (Lee et al., 2008).

For a large-scale network integration, Karr et al. (2012) collected information from 900 data sources, including reviews, books and databases, and constructed a whole cell model of *Mycoplasma genitalium*. This model includes data on metabolism, signal transduction and transcriptional regulation, and offers deep understanding on many previously unknown cellular behaviors, such as the inverse relationship between the replication rates and durations of DNA replication initiation. Furthermore, experimental analysis based on the model predictions has certified several previously undetected biological functions and kinetic parameters (Karr et al., 2012). However, due to the particularity of the species itself (e.g., unclear medium component, too small genome, etc.), the experimental data is rare, so the model was built using lots of data from other species, which makes it not suitable for other species. The good news is that Carrera et al. (2014) proposed a widely applicable modeling methodology for integrated network reconstruction and reconstructed an *E. coli* metabolic-regulatory-signaling integrated network by combining high-throughput transcriptome and phenomic data. The methodology is composed of four different algorithms including Expression Balance Analysis (EBA), flux Variability Analysis (FVA), TRAnscription-based Metabolic flux Enrichment (TRAME) and FBA, which were sequentially used to calculate the gene expression caused by the genetic or environmental perturbations, the flux balance bounds modified by the predicted gene expression, the metabolism-transcription interactions, and the optimized objective function under the modified flux bounds. With this methodology, the metabolism, transcription, and signal transduction information were integrated into one computable model. The application of this methodology on *E. coli* showed that the integrated network has a more powerful capability in phenotype prediction than the approaches using metabolic network alone (Carrera et al., 2014).

THE INTEGRATED NETWORKS OF MICROORGANISMS AND HUMAN DISEASES

As many microorganisms are closely related to non-infectious human diseases, their biological networks naturally provide a possibility for studying the complex mechanisms of human diseases. For example, signal and metabolic network are usually used to understand the mechanism of disease and drug discovery (Hasan et al., 2012). In this point of view, another type of integrated network, microbe-disease association network integrated with microorganisms and human diseases, is also a quite helpful tool for improving the treatment of human diseases or development of new drugs. Up to date some efforts has been made to develop the algorithms or models for predicting the disease-related microorganisms based on the microbe-disease association network. Chen et al. (2017) developed a computational model KATZHMDA (KATZ measure for Human Microorganism–Disease Association prediction) based on an assumption that microorganisms with similar function likely to have similar interactions and non-interactions with diseases. With the similar assumption, Huang Y.A. et al. (2017) also developed a computational model called NGRHMDA (a neighbor- and graph-based combined recommendation model for human microbe-disease association prediction) to

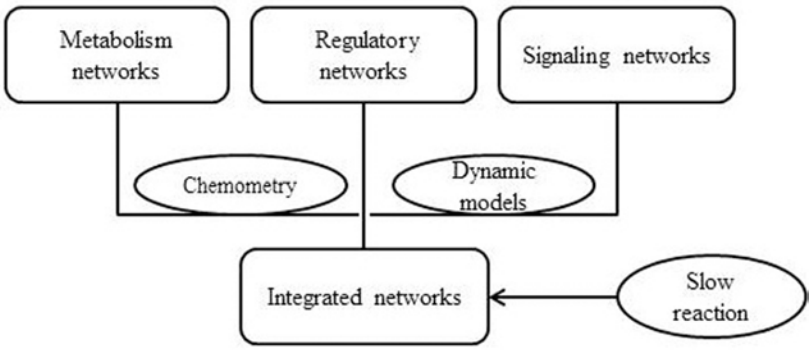


FIGURE 3 | Schematic diagram of idFBA.

predict the association between microorganisms and diseases. They used a graph-based scoring method and neighbor-based collaborative filtering to calculate the possibility of association between microorganisms and diseases (Huang Y.A. et al., 2017). Huang Z.A. et al. (2017) developed a computational model PBHMDA (Path-Based Human Microorganism-Disease Association prediction) based on the Gaussian interaction profile kernel similarity calculation for microorganisms and diseases. Besides, this model also integrated the known microbe-disease relationships, and part of the results predicted with this model has been confirmed by previous published literature (Huang Z.A. et al., 2017). Similarly, Wang F. et al. (2017) proposed a semi-supervised computational model LRLSHMDA (Laplacian Regularized Least Squares for Human Microorganism-Disease Association) by integrating the Gaussian interaction profile kernel similarity and Laplacian regularized least squares classifier. This model got good performance on the prediction of chronic obstructive pulmonary, colorectal carcinoma, and asthma diseases in the case studies (Wang F. et al., 2017). No matter what kind of algorithms, the predictions were made based on the known knowledge of microorganisms and microbe-disease relationships. Therefore, as we know more about microbes and diseases, the computational models are expected to offer more insights in the identification of microbe-disease associations in the future.

FUTURE OF MICROBE CELLULAR NETWORK

Construction and analysis of biological information processing-specific large-scale cellular networks (i.e., metabolic, signaling, and gene regulatory networks) has output many important biological insights in novel pathways, regulatory, and metabolic mechanisms. Given the fact that these networks are highly interconnected, the analysis of the integrated networks is expected to supply more novel understanding on biological behaviors which cannot be achievable using the biological information processing-specific network models alone. From biological information processing-specific networks to integrated network, it is an irresistible trend of the analysis of cellular networks. The integrated networks may provide better answers to the issues such as how transcription-regulatory interactions redirect flux distribution in a metabolic network; how a environmental or genetic disturbance influences the phenotype of an organism; or giving more accurate suggestions to the experiment designs and driving biotechnology applications. As lots of information is required in the reconstruction of a large-scale integrated networks, high-throughput experiments will play an increasingly significant role in the network integration. With the development of sequencing technology in recent years, many other types of cellular molecules involved in the regulatory process has been identified with high throughput experiment, and their related cellular networks have been studied, such as the network of mRNA, microRNA (Ferguson et al., 2018), lncRNAs (Zhang et al., 2018), and ceRNA (Xue et al., 2018). These small molecules participate in the regulatory network and control the RNA activity or gene expression directly or indirectly. Therefore, the integration of these molecules with TFs provides more information to the TRNs (Wong and Matus, 2017). With the involvement of more types of elements in the molecular networks, the integrated cellular networks will perform better to simulate the activity of the real cells. Although integrating of multiple types of information into a network will largely increase its complexity and calculation difficulties, the integrated network makes a computational network closer to a real cell, which pushes us go further from the dream of reproducing real creatures on computers.

ETHICS STATEMENT

The study was approved by College of Life Sciences, Tianjin Normal Univeristy.

AUTHOR CONTRIBUTIONS

DW, LZ, and QW collected the references. EW and JS contributed in the guideline and revision of the manuscript. TH analyzed the reference. TH and DW wrote the paper.

FUNDING

This work was supported by Grants of the Major State Basic Research Development Program of China (973 programs, 2012CB114405), National Natural Science Foundation of China (31770904, 21106095), Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC30700), Project of introducing one thousand high level talents in three years, "131" Innovative Talents cultivation of Tianjin, Academic Innovation Foundation of Tianjin Normal University (52XC1403).

REFERENCES

Arrieta-Ortiz, M. L., Hafemeister, C., Bate, A. R., Chu, T., Greenfield, A., Shuster, B., et al. (2015). An experimentally supported model of the *Bacillus subtilis* global transcriptional regulatory network. *Mol. Syst. Biol.* 11:839. doi: 10.15252/msb.20156236

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* 14, 283–291. doi: 10.1016/j.sbi.2004.05.004

Barrett, C. L., Herring, C. D., Reed, J. L., and Palsson, B. O. (2005). The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states. *Proc. Natl. Acad. Sci. U.S.A.* 102, 19103–19108. doi: 10.1073/pnas.0505231102

Brent, M. R. (2016). Past roadblocks and new opportunities in transcription factor network mapping. *Trends Genet.* 32, 736–750. doi: 10.1016/j.tig.2016.08.009

Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A., and Tagkopoulos, I. (2014). An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *Escherichia coli*. *Mol. Syst. Biol.* 10:735. doi: 10.15252/msb.20145108

Chandrasekaran, S., and Price, N. D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17845–17850. doi: 10.1073/pnas.1005139107

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/btx773

Chen, X., Zhao, M., and Qu, H. (2015). Cellular metabolic network analysis: discovering important reactions in *Treponema pallidum*. *Biomed. Res. Int.* 2015:328568. doi: 10.1155/2015/328568

Cong, J. (2010). *A Chemoinformatic Analysis on Metabolic Network and the Application for Screening of Drug Tagets on Erwinia Carotovora*. Master's thesis, Sichuan University, Chengdu.

Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92–96. doi: 10.1038/nature02456

Covert, M. W., and Palsson, B. O. (2002). Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.* 277, 28058–28064. doi: 10.1074/jbc.M201691200

Covert, M. W., Xiao, N., Chen, T. J., and Karr, J. R. (2008). Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* 24, 2044–2050. doi: 10.1093/bioinformatics/btn352

De-nan, Z. (2014). *Research on Prediction and Analysis of Transcriptional Regulation and Construction of Regulatory Networks Based on High-Throughout Seqyencing Data*. Harbin: Harbin Institute of Technology.

Deyasi, K., Banerjee, A., and Deb, B. (2015). Phylogeny of metabolic networks: a spectral graph theoretical approach. *J. Biosci.* 40, 799–808. doi: 10.1007/s12038-015-9562-0

Edwards, J. S., and Palsson, B. O. (1999). Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* 274, 17410–17416. doi: 10.1074/jbc.274.25.17410

Faria, J. P., Overbeek, R., Taylor, R. C., Conrad, N., Vonstein, V., Goelzer, A., et al. (2016). Reconstruction of the regulatory network for *Bacillus subtilis* and reconciliation with gene expression data. *Front. Microbiol.* 7:275. doi: 10.3389/fmicb.2016.00275

Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L., and Palsson, B. O. (2009). Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129–143. doi: 10.1038/nrmicro1949

Feist, A. M., and Palsson, B. O. (2008). The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat. Biotechnol.* 26, 659–667. doi: 10.1038/nbt1401

Ferguson, S. W., Wang, J., Lee, C. J., Liu, M., Neelamegham, S., Canty, J. M., et al. (2018). The microRNA regulatory landscape of MSC-derived exosomes: a systems view. *Sci. Rep.* 8:1419. doi: 10.1038/s41598-018-19581-x

Fowler, C. C., and Galan, J. E. (2018). Decoding a *Salmonella* Typhi regulatory network that controls typhoid toxin expression within human cells. *Cell Host Microbe* 23, 65–76.e6. doi: 10.1016/j.chom.2017.12.001

Francke, C., Siezen, R. J., and Teusink, B. (2005). Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.* 13, 550–558. doi: 10.1016/j.tim.2005.09.001

Freyre-Gonzalez, J. A., Manjarrez-Casas, A. M., Merino, E., Martinez-Nunez, M., Perez-Rueda, E., and Gutierrez-Rios, R. M. (2013). Lessons from the modular organization of the transcriptional regulatory network of *Bacillus subtilis*. *BMC Syst. Biol.* 7:127. doi: 10.1186/1752-0509-7-127

Goelzer, A., Bekkal Brikci, F., Martin-Verstraete, I., Noirot, P., Bessieres, P., Aymerich, S., et al. (2008). Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst. Biol.* 2:20. doi: 10.1186/1752-0509-2-20

Gomez, S. M., Lo, S. H., and Rzhetsky, A. (2001). Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics* 159, 1291–1298.

Gui, J., Li, X., Zhao, P., Liu, Z., Wang, H., and Zhang, Y. (2012). A transcriptional regularoty network of *Bacillus pumilus* predicted by comparative genomics methods. *J. Sichuan Univ.* 49, 230–238.

Hao, T., Han, B., Ma, H., Fu, J., Wang, H., Wang, Z., et al. (2013). *In silico* metabolic engineering of *Bacillus subtilis* for improved production of riboflavin, Egl-237, (R,R)-2,3-butanediol and isobutanol. *Mol. Biosyst.* 9, 2034–2044. doi: 10.1039/c3mb25568a

Hao, T., Ma, H., and Zhao, X. (2012). Progress in automatic reconstruction and analysis tools of genome-scale metabolic network. *Chin. J. Biotech.* 70:661.

Hasan, S., Bonde, B. K., Buchan, N. S., and Hall, M. D. (2012). Network analysis has diverse roles in drug discovery. *Drug Discov. Today* 17, 869–874. doi: 10.1016/j.drudis.2012.05.006

Herrero, A., Reis-Cardoso, M., Jimenez-Gomez, I., Doherty, C., Agudo-Ibanez, L., Pinto, A., et al. (2017). Characterisation of HRas local signal transduction networks using engineered site-specific exchange factors. *Small GTPases* 26, 1–13. doi: 10.1080/21541248.2017.1406434

Herrgard, M. J., Lee, B. S., Portnoy, V., and Palsson, B. O. (2006). Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res.* 16, 627–635. doi: 10.1101/gr.4083206

Hoxhaj, G., Hughes-Hallett, J., Timson, R. C., Ilagan, E., Yuan, M., Asara, J. M., et al. (2017). The mTORC1 signaling network senses changes in cellular purine nucleotide levels. *Cell Rep.* 21, 1331–1346. doi: 10.1016/j.celrep.2017.10.029

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

Huang, Z. A., Chen, X., Zhu, Z., Liu, H., Yan, G. Y., You, Z. H., et al. (2017). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Ignatius Pang, C. N., Goel, A., and Wilkins, M. R. (2018). Investigating the network basis of negative genetic interactions in *Saccharomyces cerevisiae* with integrated biological networks and triplet motif analysis. *J. Proteome Res.* doi: 10.1021/acs.jproteome.7b00649 [Epub ahead of print].

Imam, S., Schauble, S., Brooks, A. N., Baliga, N. S., and Price, N. D. (2015). Data-driven integration of genome-scale regulatory and metabolic network models. *Front. Microbiol.* 6:409. doi: 10.3389/fmicb.2015.00409

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356. doi: 10.1016/S0022-2836(61)80072-7

Jiang, J. G., Song, L. F., and Zheng, P. (2012). Construction and structural analysis of integrated cellular network of *Corynebacterium glutamicum*. *Chin. J. Biotech.* 28, 577–591.

Kamps, D., and Dehmelt, L. (2017). Deblurring signal network dynamics. *ACS Chem. Biol.* 12, 2231–2239. doi: 10.1021/acschembio.7b00451

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., et al. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401. doi: 10.1016/j.cell.2012.05.044

Kashaf, S. S., Angione, C., and Lio, P. (2017). Making life difficult for *Clostridium difficile*: augmenting the pathogen's metabolic model with transcriptomic and codon usage data for better therapeutic target characterization. *BMC Syst. Biol.* 11:25. doi: 10.1186/s12918-017-0395-3

Kim, B., Kim, W. J., Kim, D. I., and Lee, S. Y. (2015). Applications of genome-scale metabolic network model in metabolic engineering. *J. Ind. Microbiol. Biotechnol.* 42, 339–348. doi: 10.1007/s10295-014-1554-9

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492

Kromer, J. O., Sorgenfrei, O., Klopprogge, K., Heinzle, E., and Wittmann, C. (2004). In-depth profiling of lysine-producing *Corynebacterium glutamicum* by combined analysis of the transcriptome, metabolome, and fluxome. *J. Bacteriol.* 186, 1769–1784. doi: 10.1128/JB.186.6.1769-1784.2004

Le Novere, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., et al. (2009). The systems biology graphical notation. *Nat. Biotechnol.* 27, 735–741. doi: 10.1038/nbt.1558

Lee, J. M., Gianchandani, E. P., Eddy, J. A., and Papin, J. A. (2008). Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput. Biol.* 4:e1000086. doi: 10.1371/journal.pcbi.1000086

Levchenko, A. (2003). Dynamical and integrative cell signaling: challenges for the new biology. *Biotechnol. Bioeng.* 84, 773–782. doi: 10.1002/bit.10854

Levering, J., Dupont, C. L., Allen, A. E., Palsson, B. O., and Zengler, K. (2017). Integrated regulatory and metabolic networks of the marine diatom *Phaeodactylum tricornutum* predict the response to rising CO_2 levels. *mSystems* 2:e00142-16. doi: 10.1128/mSystems.00142-16

Liu, W., Li, D., Zhu, Y., and He, C. (2008). Bioinformatics analyses for signal transduction networks. *Sci. China C Life Sci.* 11, 994–1002. doi: 10.1007/s11427-008-0134-5

Ma, H. W., and Zeng, A. P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 19, 1423–1430. doi: 10.1093/bioinformatics/btg177

Mahadevan, R., Bond, D. R., Butler, J. E., Esteve-Nunez, A., Coppi, M. V., Palsson, B. O., et al. (2006). Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. *Appl. Environ. Microbiol.* 72, 1558–1568. doi: 10.1128/AEM.72.2.1558-1568.2006

Notebaart, R. A., van Enckevort, F. H., Francke, C., Siezen, R. J., and Teusink, B. (2006). Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* 7:296. doi: 10.1186/1471-2105-7-296

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., et al. (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Mol. Syst. Biol.* 7:535. doi: 10.1038/msb.2011.65

Pal, C., Papp, B., Lercher, M. J., Csermely, P., Oliver, S. G., and Hurst, L. D. (2006). Chance and necessity in the evolution of minimal metabolic networks. *Nature* 440, 667–670. doi: 10.1038/nature04568

Papin, J. A., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* 6, 99–111. doi: 10.1038/nrm1570

Papin, J. A., and Palsson, B. O. (2004). The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J.* 87, 37–46. doi: 10.1529/biophysj.103.029884

Piubeli, F., Salvador, M., Argandona, M., Nieto, J. J., Bernal, V., Pastor, J. M., et al. (2018). Insights into metabolic osmoadaptation of the ectoines-producer bacterium Chromohalobacter salexigens through a high-quality genome scale metabolic model. *Microb. Cell Fact.* 17:2. doi: 10.1186/s12934-017-0852-0

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi: 10.1126/science.1073374

Rother, M., Munzner, U., Thieme, S., and Krantz, M. (2013). Information content and scalability in signal transduction network reconstruction formats. *Mol. Biosyst.* 9, 1993–2004. doi: 10.1039/c3mb00005b

Roy, S., Lagree, S., Hou, Z., Thomson, J. A., Stewart, R., and Gasch, A. P. (2013). Integrated module and gene-specific regulatory inference implicates upstream signaling networks. *PLoS Comput. Biol.* 9:e1003252. doi: 10.1371/journal.pcbi.1003252

Ryll, A., Bucher, J., Bonin, A., Bongard, S., Goncalves, E., Saez-Rodriguez, J., et al. (2014). A model integration approach linking signalling and gene-regulatory logic with kinetic metabolic models. *Biosystems* 124, 26–38. doi: 10.1016/j.biosystems.2014.07.002

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. (2002). Genome-scale metabolic model of *Helicobacter* pylori 26695. *J. Bacteriol.* 184, 4582–4593. doi: 10.1128/JB.184.16.4582-4593.2002

Shlomi, T., Eisenberg, Y., Sharan, R., and Ruppin, E. (2007). A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol. Syst. Biol.* 3:101. doi: 10.1038/msb4100141

Shlomi, T., Segal, D., Ruppin, E., and Sharan, R. (2006). QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics* 7:199. doi: 10.1186/1471-2105-7-199

Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. (2008). DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res.* 36, D93–D96. doi: 10.1093/nar/gkm910

Simeonidis, E., Chandrasekaran, S., and Price, N. D. (2013). A guide to integrating transcriptional regulatory and metabolic networks using PROM (probabilistic regulation of metabolism). *Methods Mol. Biol.* 985, 103–112. doi: 10.1007/978-1-62703-299-5_6

Soh, K. C., and Hatzimanikatis, V. (2010). DREAMS of metabolism. *Trends Biotechnol.* 28, 501–508. doi: 10.1016/j.tibtech.2010.07.002

Steinway, S. N., Biggs, M. B., Loughran, T. P. Jr., Papin, J. A., and Albert, R. (2015). Inference of network dynamics and metabolic interactions in the gut microbiome. *PLoS Comput. Biol.* 11:e1004338. doi: 10.1371/journal.pcbi.1004338

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420, 190–193. doi: 10.1038/nature01166

Thiele, I., and Palsson, B. O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121. doi: 10.1038/nprot.2009.203

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., et al. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* 4:rs8. doi: 10.1126/scisignal.2001699

Wang, E., Lenferink, A., and O'Connor-McCourt, M. (2007). Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cell Mol. Life Sci.* 64, 1752–1762. doi: 10.1007/s00018-007-7054-6

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., et al. (2017). LRLSHMDA: laplacian regularized least squares for human microbe-disease association prediction. *Sci. Rep.* 7:7601. doi: 10.1038/s41598-017-08 1298127-2

Wang, H., Ma, H., and Zhao, X. (2010). Progress in genome-scale metabolic network: a review. *Chin. J. Biotech.* 26, 1340–1348.

Wang, K., Hu, F., Xu, K., Cheng, H., Jiang, M., Feng, R., et al. (2011). CASCADE_SCAN: mining signal transduction network from high-throughput data based on steepest descent method. *BMC Bioinformatics* 12:164. doi: 10.1186/1471-2105-12-164

Wang, R., Su, C., Wang, X., Fu, Q., Gao, X., Zhang, C., et al. (2018). Global gene expression analysis combined with a genomics approach for the identification of signal transduction networks involved in postnatal mouse myocardial

proliferation and development. *Int. J. Mol. Med.* 41, 311–321. doi: 10.3892/ijmm.2017.3234

Wang, Y. C., and Chen, B. S. (2010). Integrated cellular network of transcription regulations and protein-protein interactions. *BMC Syst. Biol.* 4:20. doi: 10.1186/1752-0509-4-20

Wang, Z., Danziger, S. A., Heavner, B. D., Ma, S., Smith, J. J., Li, S., et al. (2017). Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast. *PLoS Comput. Biol.* 13:e1005489. doi: 10.1371/journal.pcbi.1005489

Wong, D. C. J., and Matus, J. T. (2017). Constructing integrated networks for identifying new secondary metabolic pathway regulators in grapevine: recent applications and future opportunities. *Front. Plant Sci.* 8:505. doi: 10.3389/fpls.2017.00505

Wu, M., and Ma, H. (2014). The progress of integrated genome-scale cellular networks. *Microbiol. China* 41, 367–375.

Xue, W. H., Fan, Z. R., Li, L. F., Lu, J. L., Ma, B. J., Kan, Q. C., et al. (2018). Construction of an oesophageal cancer-specific ceRNA network based on miRNA, lncRNA, and mRNA expression data. *World J. Gastroenterol.* 24, 23–34. doi: 10.3748/wjg.v24.i1.23

Yilmaz, L. S., and Walhout, A. J. (2017). Metabolic network modeling with model organisms. *Curr. Opin. Chem. Biol* 36, 32–39. doi: 10.1016/j.cbpa.2016.12.025

Zhang, Y., Tao, Y., Li, Y., Zhao, J., Zhang, L., Zhang, X., et al. (2018). The regulatory network analysis of long noncoding RNAs in human colorectal cancer. *Funct. Integr. Genomics* doi: 10.1007/s10142-017-0588-2 [Epub ahead of print].

Zheng, Y., Du, S., Xu, H., Xu, Y., Zhao, H., Chi, T., et al. (2014). [Cellular adhesion signal transduction network of tumor necrosis factor-alpha induced hepatocellular carcinoma cells]. *Zhonghua Yi Xue Za Zhi* 94, 3345–3348.

Zhao, J. (2008). *Research on Structure,Function and Evolution of Cellur Metabolic Networks*. Master's thesis, Shanghai Jiaotong University, Shanghai.

Conflict of Interest Statement: 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.

Copyright © 2018 Hao, Wu, Zhao, Wang, Wang and Sun. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 16 March 2018
doi: 10.3389/fmicb.2018.00476

PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine

Balachandran Manavalan[1], *Tae H. Shin*[1,2] *and Gwang Lee*[1,2]*

[1] Department of Physiology, Ajou University School of Medicine, Suwon, South Korea, [2] Institute of Molecular Science and Technology, Ajou University, Suwon, South Korea

Accurately identifying bacteriophage virion proteins from uncharacterized sequences is important to understand interactions between the phage and its host bacteria in order to develop new antibacterial drugs. However, identification of such proteins using experimental techniques is expensive and often time consuming; hence, development of an efficient computational algorithm for the prediction of phage virion proteins (PVPs) prior to *in vitro* experimentation is needed. Here, we describe a support vector machine (SVM)-based PVP predictor, called PVP-SVM, which was trained with 136 optimal features. A feature selection protocol was employed to identify the optimal features from a large set that included amino acid composition, dipeptide composition, atomic composition, physicochemical properties, and chain-transition-distribution. PVP-SVM achieved an accuracy of 0.870 during leave-one-out cross-validation, which was 6% higher than control SVM predictors trained with all features, indicating the efficiency of the feature selection method. Furthermore, PVP-SVM displayed superior performance compared to the currently available method, PVPred, and two other machine-learning methods developed in this study when objectively evaluated with an independent dataset. For the convenience of the scientific community, a user-friendly and publicly accessible web server has been established at www.thegleelab.org/PVP-SVM/PVP-SVM.html.

Keywords: bacteriophage virion proteins, feature selection, hybrid features, machine learning, support vector machine

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Wei Chen,
North China University of Science and Technology, China

***Correspondence:**
Gwang Lee
glee@ajou.ac.kr

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

Received: *07 December 2017*
Accepted: *28 February 2018*
Published: *16 March 2018*

Citation:
Manavalan B, Shin TH and Lee G (2018) PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. Front. Microbiol. 9:476. doi: 10.3389/fmicb.2018.00476

INTRODUCTION

Bacteriophages, also known as phages, are viruses that can infect and replicate in bacteria, and are found wherever bacteria survive. The phage virion is composed of proteins that encapsulate either DNA or RNA, which binds to bacterial surface and injects its genetic materials into the specific host bacteria. In lytic cycle, phage genes are expressed for proteins that poke hole in the cell membrane, which makes cell expand and burst. Subsequently, released phages from cell bursting spread and infects other host cells. Identification of phage virion proteins (PVPs) is important for understanding the relationship between phage and host bacteria and also development of novel antibacterial drugs or antibiotics (Lekunberri et al., 2017). For instance, phage encoded proteins including endolysins, exopolysaccharidases, and holins have been proven as promising antibacterial products (Drulis-Kawa et al., 2012). Experimental methods, including mass spectrometry, sodium

dodecyl sulfate polyacrylamide gel electrophoresis, and protein arrays (Lavigne et al., 2009; Yuan and Gao, 2016; Jara-Acevedo et al., 2018) have been used to identify PVPs. However, these methods are expensive and often time-consuming. Therefore, computational methods to predict PVPs prior to *in vitro* experimentation are needed. It is difficult to predict the function of PVPs from sequence information because of relatively limited experimental data. However, machine-learning (ML) approaches have been successfully applied to several similar biological problems. Therefore, it may be possible to predict the functions of phage proteins using ML.

To this end, Seguritan et al., developed the first method to classify viral structure proteins using an artificial neural network, using amino acid composition (AAC) and protein isoelectric points as input features (Seguritan et al., 2012). Later, Feng et al., developed a naïve Bayesian method, with an algorithm utilizing AAC and dipeptide composition (DPC) as input features (Feng et al., 2013b). Subsequently, Ding et al., developed a support vector machine (SVM)-based prediction model called PVPred. In this method, analysis of variance was applied to select important features from g-gap DPC (Ding et al., 2014). Recently, Zhang et al., developed a random forest (RF)-based ensemble method to distinguish PVPs and non-PVPs (Zhang et al., 2015). PVPred is the only existing publicly available method that was developed using the same dataset as our method. Although the existing methods have specific advantages in PVPs prediction, it remains necessary to improve the accuracy and transferability of the prediction model.

It is worth mentioning that several sequence-based features including AAC, atomic composition (ATC), chain-transition-distribution (CTD), DPC, pseudo amino acid composition and amino acid pair, and several feature selection techniques including correlation-based feature selection, ANOVA feature selection, minimum-redundancy and maximum-relevance, RF-algorithm based feature selection have been successfully applied in other protein bioinformatics studies (Wang et al., 2012, 2016; Lin et al., 2015; Qiu et al., 2016; Tang et al., 2016; Gupta et al., 2017; Manavalan and Lee, 2017; Manavalan et al., 2017; Song et al., 2017). All these studies motivated us in the development of a new model in this study. Hence, we developed a SVM-based PVP predictor called PVP-SVM, in which the optimal features were selected using a feature selection protocol that has been successfully applied to various biological problems (Manavalan and Lee, 2017). We selected the optimal features from a large set, including AAC, DPC, CTD, ATC, and PCP. In addition to SVM (i.e., PVP-SVM), we also developed RF and extremely randomized tree (ERT)-based methods. The performance of PVP-SVM was consistent in both the training and independent datasets, and was superior to the current method and the RF and ERT methods developed in this study.

MATERIALS AND METHODS

Training Dataset

In this study, we utilized the dataset constructed by Ding et al., which was specifically used for studying PVPs (Ding et al., 2014). We decided to use this dataset for the following reasons: (i) it is a reliable dataset, constructed based on several filtering schemes; (ii) it is a non-redundant dataset and none of the sequences possesses pairwise sequence identity (>40%) with any other sequence. Hence, this dataset stringently excludes homologous sequences; and (iii) most importantly, it facilitates fair comparison between the current method and existing methods, which were developed using the same training dataset. Thus, the training dataset can be formulated as:

$$S = S^{+} \cup S^{-} \tag{1}$$

where the positive subset S^{+} contained 99 PVPs, the negative subset S^{-} contained 208 non-PVPs, and the symbol ∪ denotes union in the set theory. Thus, S contained 307 samples.

Independent Dataset

We obtained PVP and non-PVP sequences from the Universal Protein Resource (UniProt) as previously described (Feng et al., 2013b; Ding et al., 2014; Zhang et al., 2015). To avoid overestimation in the prediction model, we excluded sequences that shared greater than 40% sequence identity with sequences in the training dataset. The final dataset contained 30 PVPs and 64 non-PVPs. We note that our independent dataset included Ding et al., independent dataset. The above two datasets can be downloaded from our web server.

Input Features

(i) AAC: The fractions of the 20 naturally occurring amino acid residues in a given protein sequence were calculated as follows:

$$\mathrm{AAC}\,(i) = \frac{Frequency\ of\ amino\ acid\ (i)}{Length\ of\ the\ protein\ sequence} \tag{2}$$

where i can be any of the 20 natural amino acids.

(ii) ATC: The fraction of five atom types (C, H, N, O, and S) in a given protein sequence was calculated as previously reported (Kumar et al., 2015; Manavalan et al., 2017), with a fixed length of five features.

(iii) CTD: The global composition feature encoding method CTD comprises properties such as hydrophobicity, polarity, normalized van der Waals volume, polarizability, predicted secondary structure, and solvent accessibility. It was first proposed in protein folding class prediction (Dubchak et al., 1995). Composition (C) represents the composition percentage of each group in the peptide sequence. Transition (T) represents the transition probability between two neighboring amino acids belonging to two different groups. Distribution (D) represents the position of amino acids (the first 25, 50, 75, or 100%) in each group in the protein sequence. For each qualitative property of a given sequence, C, T, and D produce 3, 3, and 15-dimension features, respectively. As a result, $7 \times (3 + 3 + 15) = 147$ features can be generated for seven qualitative properties.

(iv) DPC: The fractions of the 400 possible dipeptides present in a given protein sequence were calculated as follows:

$$\mathrm{DPC}(j) = \frac{Total\ number\ of\ dipeptide\ (j)}{Total\ number\ of\ all\ possible\ dipeptides} \tag{3}$$

where *j* can be any of the 400 possible dipeptides.

(v) PCP: We employed 11 representative PCP attributes of amino acids for feature extraction (polar, hydrophobic, charged, aliphatic, aromatic, positively charged, negatively charged, small, tiny, large, and peptide mass).

Note that all of the above features were in the range of [0, 1] as input for training and testing.

The Support Vector Machine

We employed a SVM as our classification algorithm, a well-known supervised ML method introduced in Boser et al. (1992) that has been applied to several biological problems (Wang et al., 2009; Eickholt et al., 2011; Deng et al., 2013; Cao et al., 2014; Manavalan et al., 2015). The objective of a SVM is to find the hyperplane with the largest margin to decrease the misclassification rate. Given a set of data points (input features) and an objective function associated with the data points (PVPs: 1 and non-PVPs: 0), SVM learn a function in the form of

$$y = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \quad (4)$$

where y is the predicted class associated with an input feature vector of x; α_i is the adjustable weight assigned to the training data point x_i during training by minimizing a quadratic objective function; b is the bias term; and K is the Kernel function. Therefore, y can be viewed as a weighted linear combination of similarities between the training data points x_i and the target data point x. Data points with positive weights in the training dataset affect the final solution and are called support vectors. SVM is especially effective when the input data are not linearly separable. K is required to map the input data into a higher dimensional space to identify the optimal separating hyperplane (Scholkopf and Smola, 2001). Therefore, we experimented with several common Ks, including linear, Gaussian radial basis, and polynomial functions. The Gaussian radial basis K ($e^{(-\gamma \times \|x-y\|^2)}$; $\gamma = \frac{1}{\sigma^2}$) performed the best. Here, two critical parameters (γ and C) required optimization: γ controls how peaked Gaussians are centered on the support vectors, while C controls the trade-off between the training error and the margin size (Smola and Vapnik, 1997; Vapnik and Vapnik, 1998; Scholkopf and Smola, 2001). These two parameters were optimized using a grid search from 2^{-15}–2^{10} for C and 2^{-10}–2^{10} for γ, in $\log_2$ steps. In this study, we used a SVM implemented in the scikit-learn package (Pedregosa et al., 2011).

Cross-Validation and Independent Testing

As demonstrated in a series of studies (Feng et al., 2013a,c, 2018; Chen et al., 2014, 2017a,b), among three cross-validation methods, i.e., independent dataset test, K-fold cross-validation test and Leave-one-out cross-validation (LOOCV, also called jackknife cross validation), LOOCV is the most rigorous and objective evaluation methods. Accordingly, the jackknife test has been widely recognized and increasingly used to test the quality for various predictors. In LOOCV, each sample in the training dataset is in turn singled out as an independent test sample and all the rule parameters are calculated without including the one being identified. We performed LOOCV on the training dataset and the trained model was tested on the independent dataset to confirm the generality of the developed method.

Performance Evaluation Criteria

The following four metrics are commonly used in literature to measure the quality of binary classification (Xiong et al., 2012; Li et al., 2015): sensitivity, specificity, accuracy and Matthews' correlation coefficient (MCC), which are expressed as

$$\begin{cases} Sensitivity = \frac{TP}{TP+FN} \\ Specificity = \frac{TN}{TN+FP} \\ Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \quad (5)$$

where TP is the number of PVPs predicted to be PVPs; TN is number of non-PVPs predicted to be non-PVP; FP is the number of non-PVPs predicted to be PVP; and FN is the number of PVPs predicted to be non-PVP.

To further evaluate the performance of the classifier, we employed a receiver operating characteristic (ROC) curve. The ROC curve was plotted with the false positive rate as the x-axis and true positive rate as the y-axis by varying the thresholds. The area under the curve (AUC) was used for model evaluation, with higher AUC values corresponding to better performance of the classifier.

RESULTS

Framework of the Proposed Predictor

Figure 1 illustrates the overall framework of the PVP-SVM method. It consisted of four steps: (i) construction of the training and independent datasets; (ii) extraction of various features from the primary sequences, including AAC, ATC, CTD, DPC, and PCP; (iii) generation of 25 different feature sets based on feature importance scores (FIS) computed using the RF algorithm. These different sets were inputted to the SVM to develop their respective prediction models; and (iv) the model producing the best performance in terms of MCC was considered the final model, and the corresponding feature set was considered the optimal feature set.

Feature Selection Protocol

Generally, high dimensional features can contain a higher degree of irrelevant and redundant information that may greatly degrade the performance of ML algorithms. Therefore, it is necessary to apply a feature selection protocol to filter the redundant features and increase prediction efficiency (Wang et al., 2012; Zheng et al., 2012; Manavalan et al., 2014; Manavalan and Lee, 2017; Song et al., 2017). Previously, Manavalan and Lee applied a systematic feature selection protocol and developed a novel quality assessment method called SVMQA (Manavalan and Lee, 2017), which was the best method in CASP12 blind prediction experiments (Elofsson et al., 2017; Kryshtafovych et al., 2017). We applied a similar protocol in our recent studies, including cell-penetrating peptide

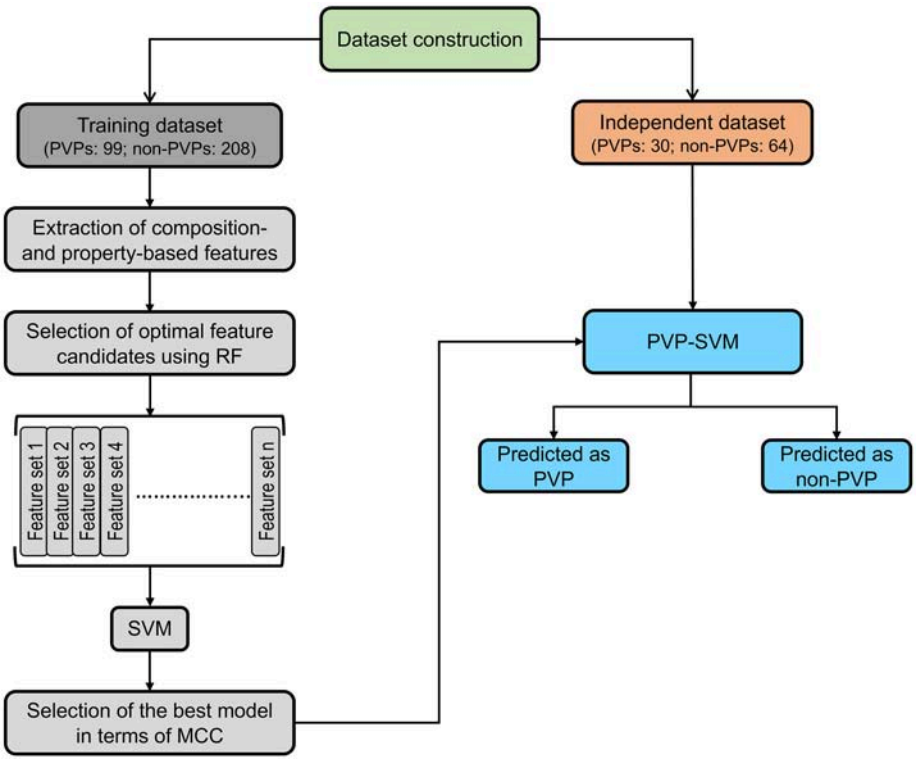


FIGURE 1 | PVP-SVM development consisted of four steps: (i) dataset construction; (ii) feature extraction; (iii) development of the prediction model; and (iv) selection of the best model and construction of PVP-SVM.

and DNase I hypersensitivity predictions (Manavalan et al., 2018). Interestingly, this protocol significantly improved the performance of our method. Therefore, we extended this approach to the current problem. The current protocol differs slightly from the published protocol in terms of parameters (*ntree* and *mtry*) used in the RF algorithm, which is mainly due to the large number of features used in this study (i.e., 26-fold more features than were used in SVMQA).

In our study, each protein sequence was represented as 583 dimensional vectors, which was higher than the number of samples. In the first step, we applied the RF algorithm and estimated the FIS of 583 features (AAC: 20; DPC: 400; ATC: 5; PCP: 11; and CTD: 147) to distinguish PVPs and non-PVPs. A detailed description of how we computed the FIS scores of the input features has been reported previously (Manavalan et al., 2014; Manavalan and Lee, 2017). Briefly, we used all features as inputs in the RF algorithm and performed ten-fold cross-validation using the training dataset. For each round of cross-validation, we built 5,000 trees, and the number of variables at each node was chosen randomly from 1 to 100. The average FIS from all the trees are shown in **Figure 2A**, where most of the features had similar scores and only ∼5% (FIS $\geq$ 0.005) contributed significantly to PVP prediction. In the second step, we applied a FIS cutoff $\geq$ 0.001 and selected 477 features as optimal feature candidates (**Figure 2B**). Subsequently, we generated 25 different sets of features from the optimal feature candidates based on an FIS cut-off (0.001 $\leq$ FIS $\leq$ 0.004, with a step size of 0.0011). Basically, we considered each set of more important features in a step-wise manner. To identify the optimal feature set, we inputted each set into the SVM separately and performed LOOCV to evaluate their performance. The prediction model that produced the best performance (i.e., the highest MCC) was considered final, and the corresponding feature set was considered optimal.

Performance of Various Prediction Models on the Training Dataset

Figure 3A shows the performances of the SVM model using different sets of input features, in which the MCC gradually increased with respect to the different feature sets, peaked with the F136-based model, and then gradually declined. **Figure 3B** shows the classification accuracy vs. parameter variation (C and γ) of the final F136-based model. The maximal classification accuracy was 0.870, when the parameters $\log_2(C)$ and $\log_2(\gamma)$ were 6.72 and −2.18, respectively, with MCC, sensitivity, and specificity values of 0.695, 0.737, and 0.933, respectively. The feature type distribution of the optimal feature set and the total features employed in this study are shown in **Figure 3C**. Among 136 optimal features, there were eight AAC features, one ATC feature, 25 CTD features, 98 DPC features, and four PCP features, indicating that important properties from all five compositions contributed to PVP prediction.

To demonstrate the effect of our feature selection protocol, we compared the F136-based model with the

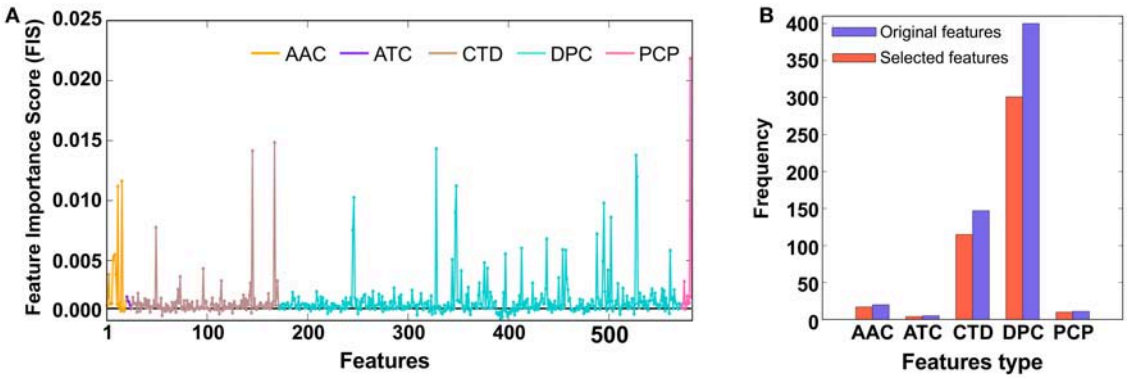


FIGURE 2 | (A) The x- and y-axes represent each feature and its feature importance scores (FIS), respectively. We applied a FIS cutoff $\geq$ 0.001 and selected 477 optimal feature candidates. **(B)** Distribution of each feature type in the optimal feature candidates and original feature set.

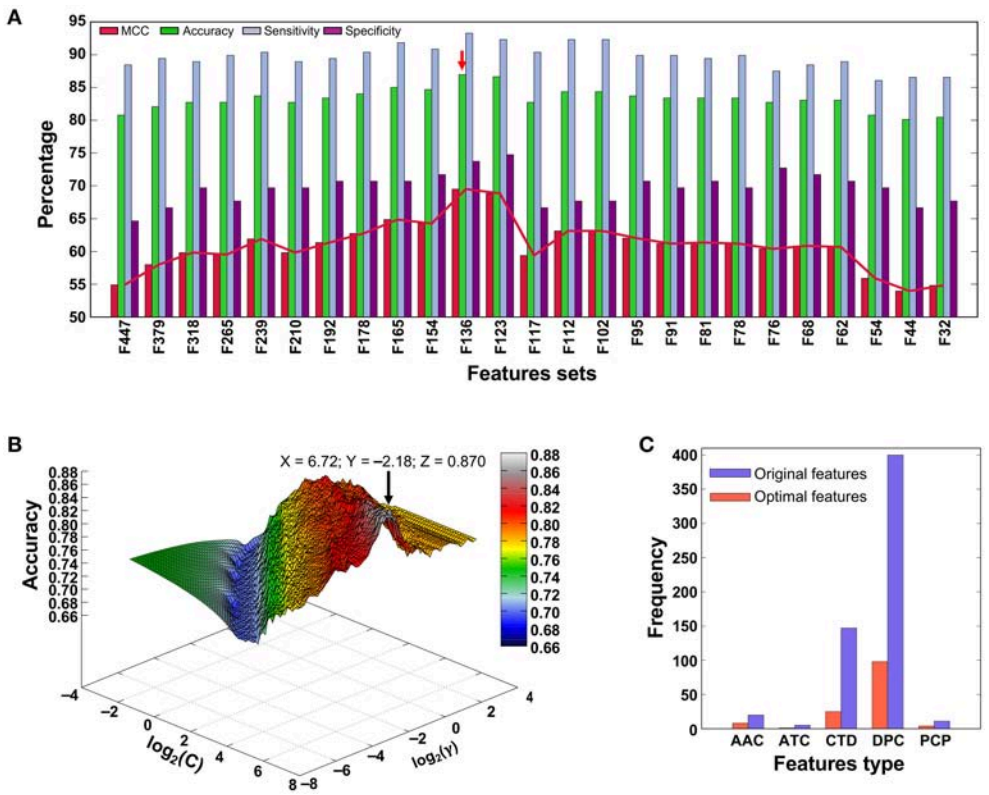


FIGURE 3 | (A) Performance of SVM-based classifiers in distinguishing between PVPs and non-PVPs. A total of 25 classifiers were evaluated using LOOCV, and their performances in terms of MCC, accuracy, sensitivity, and specificity are shown. The red arrow denotes the final selected model. **(B)** Classification accuracy of the F136-based (final selected) model with respect to variations in parameters C and γ. **(C)** Distribution of each feature type in the optimal feature set (136 features) and original feature set (583 features).

control SVM (using all features) and also an individual composition-based prediction model. As shown in **Table 1**, F136-based model accuracy, MCC, and area under curve (AUC) were 15–44, 6–17, and 6–11% higher, respectively, than the other models. These results demonstrate that the many redundant or uninformative features present in the original feature set were eliminated through our feature selection protocol, resulting in significant performance improvement.

Comparison of PVP-SVM With Other ML Algorithms

In addition to PVP-SVM, we also developed RF- and ERT-based models using the same feature selection protocol and training

dataset (**Figures 4A,B**). These two methods have been described in detail in our previous study (Manavalan et al., 2017, 2018). The procedure for ML parameter optimization and final model selection was the same as for PVP-SVM. The performance of the final selected RF and ERT models was compared with PVP-SVM, as well as PVPred, which was constructed using the same training dataset. **Table 2** shows that the accuracy, AUC, and MCC of PVP-SVM were 2–4, 0.1–2, and 8–9% higher, respectively, than those achieved by other methods, indicating the superiority of PVP-SVM.

Method Performance Using an Independent Dataset

We evaluated the performance of our three ML methods and PVPred using an independent dataset. **Table 3** shows that PVP-SVM achieved the highest MCC and AUC values (0.531 and 0.844, respectively). Indeed, the corresponding metrics were 2.2–17.4% and 4.8–10.0% higher than those achieved by other methods, indicating the superiority of PVP-SVM. Specifically, PVP-SVM outperformed PVPred in all five metrics,

TABLE 1 | A comparison of the proposed predictor with the individual composition-based SVM model on training dataset.

Methods	MCC	Accuracy	Sensitivity	Specificity	AUC	*P*-value
PVP-SVM	0.695	0.870	0.737	0.933	0.900	
SVM control	0.554	0.811	0.636	0.894	0.837	0.068
AAC	0.525	0.792	0.841	0.687	0.841	0.086
DPC	0.395	0.743	0.837	0.546	0.760	***0.00023***
CTD	0.534	0.801	0.880	0.636	0.819	***0.022***
DPC	0.478	0.782	0.889	0.556	0.812	***0.014***
ATC	0.252	0.708	0.091	1.000	0.788	***0.002***

The first column represents the method name employed in this study. The second, the third, the fourth and the fifth respectively represent the MCC, accuracy, sensitivity, and specificity. The sixth column and the seventh represent the AUC and pairwise comparison of ROC area under curves (AUCs) between PVP-SVM and the other methods using a two-tailed t-test. A $P \leq 0.05$ indicates a statistically meaningful difference between PVP-SVM and the selected method (shown in bold italic).

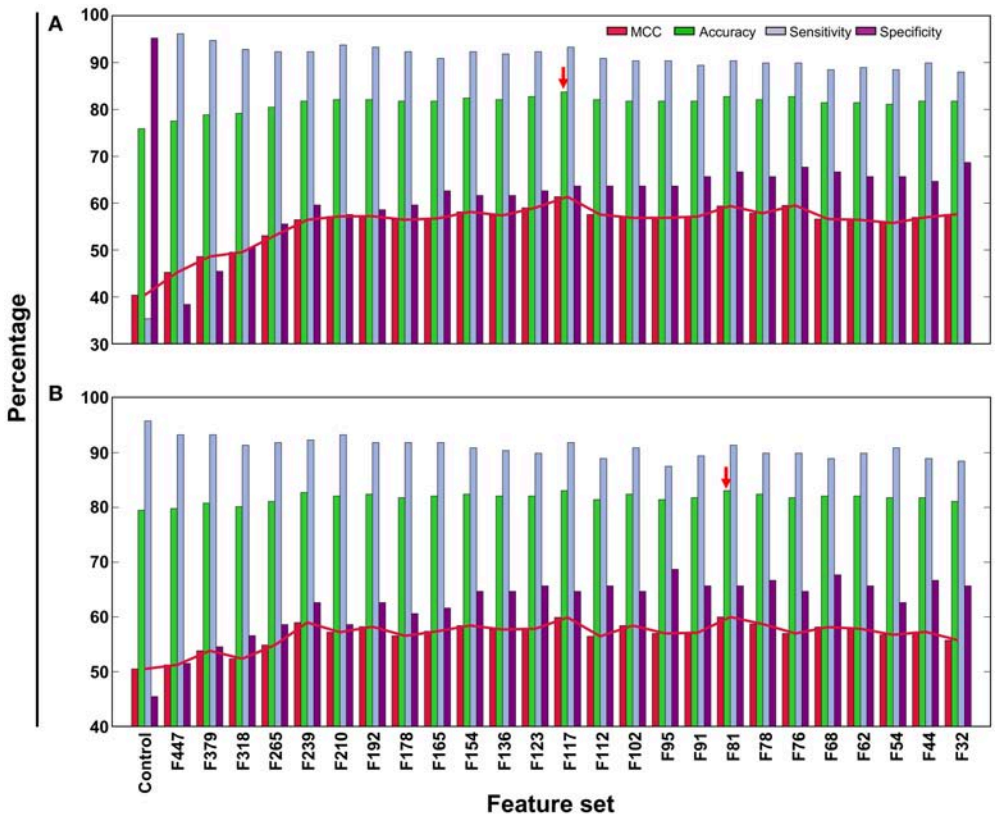


FIGURE 4 | Performance of ERT- and RF-based classifiers in distinguishing between PVPs and non-PVPs. A total of 26 classifiers were evaluated using LOOCV, whose performances in terms of MCC, accuracy, sensitivity, and specificity are shown. **(A)** ERT-based performance, **(B)** RF-based performance. Red arrow denotes the final selected models for each ML method.

TABLE 2 | A comparison of the proposed predictor with other ML-based methods on training dataset.

Methods	MCC	ACC	Sensitivity	Specificity	AUC	*P*-value
PVP-SVM	0.695	0.870	0.737	0.933	0.900	
PVPred	NA	0.850	0.758	0.894	0.899	0.974
RF	0.600	0.831	0.657	0.914	0.877	0.476
ERT	0.614	0.837	0.636	0.933	0.883	0.594

The first column represents the method name employed in this study. The second, the third, the fourth and the fifth respectively represent the MCC, accuracy, sensitivity, and specificity. The sixth column and the seventh represent the AUC and pairwise comparison of ROC area under curves (AUCs) between PVP-SVM and the other methods using a two-tailed t-test.

TABLE 3 | Performance of various methods on independent dataset.

Method	MCC	ACC	Sensitivity	Specificity	AUC	*P*-value
PVP-SVM	0.531	0.798	0.667	0.859	0.844	
ERT	0.509	0.798	0.533	0.922	0.778	0.367
RF	0.481	0.787	0.500	0.922	0.756	0.238
SVM control	0.414	0.755	0.533	0.859	0.796	0.505
PVPred	0.357	0.713	0.600	0.765	0.742	0.176

The first column represents the method name employed in this study. The second, the third, the fourth and the fifth respectively represent the MCC, accuracy, sensitivity, and specificity. The sixth column and the seventh represent the AUC and pairwise comparison of ROC area under curves (AUCs) between PVP-SVM and the other methods using a two-tailed t-test.

suggesting its usefulness as an improvement to existing tools for predicting PVPs.

In general, ML-based methods are problem-specific (Zhang and Tsai, 2005). Instead of selecting a ML method arbitrarily, it is necessary to explore different ML methods on the same dataset to select the best one. Hence, we explored three most commonly used ML methods (SVM, RF, and ERT), each having its own advantages and disadvantages. The PVP-SVM method performed consistently better than other two methods both with the training and independent datasets (**Figures 5A,B**). Although the differences in performance between these three methods were not significant ($P > 0.05$), SVM was superior to other ML methods in PVP prediction, consistent with a previous report (Ding et al., 2014). Hence, we selected PVP-SVM as the final prediction model.

Comparison of PVP-SVM and PVPred Methodology

A detailed comparison between our method and the existing method in terms of methodology is as follows: (i) the PVPred method utilizes only g-gap dipeptides as input features, and its optimal features were determined by an analysis of variance-based feature selection protocol. However, PVP-SVM utilizes AAC, ATC, CTD, and PCP in addition to DPC, with optimal features selected based on a RF algorithm; (ii) the number of optimal features used differs between the two methods; PVP-SVM uses 136 features, while PVPred uses 160; (iii) although the same ML method was used for the two methods, the parameter optimization procedure differed, as PVP-SVM used LOOCV, while PVPred used five-fold cross-validation.

Web Server Implementation

Several examples of bioinformatics tools/web servers utilized for protein function predictions have been reported in previous publications (Govindaraj et al., 2010, 2011; Manavalan et al., 2010a,b, 2011; Basith et al., 2011, 2013), and are of great practical use to researchers. To this end, an online prediction server for PVP-SVM was developed, which is freely accessible at the following link: www.thegleelab.org/PVP-SVM/PVP-SVM.html. Users can paste or upload query protein sequences in FASTA format. After submitting the input protein sequences, the results can be retrieved in a separate interface. All the curated datasets used in this study can be downloaded from the web server. PVP-SVM represents the second publicly available method for PVP prediction, and delivers a higher level of accuracy than PVPred.

DISCUSSION

PVPs play critical roles in adsorption between phages and their host bacteria, and are key in the development of new antibiotics. Phage-derived proteins are considered as safe and efficient antimicrobial agents due to its versatile properties, including bacteria-specific lytic mechanism, broad range of antibacterial spectrum, enhanced tissue penetration by small size, low immunogenicity, and reduced possibility for bacterial resistance (Drulis-Kawa et al., 2012). Thus, we have developed a novel computational method for predicting PVPs, called PVP-SVM. The molecular functions and biological activities of proteins can be predicted from their primary sequence (Lee et al., 2007); hence, we utilized the available PVPs sequences to develop the method.

A combination of AAC, ATC, DPC, CTD, and PCP features was used to map the protein sequences onto numeric feature vectors, which were inputted into the SVM to predict PVPs. Although AAC, CTD, and DPC features have been used previously (Feng et al., 2013b; Ding et al., 2014; Zhang et al., 2015), this is the first report including ATC and PCP. In ML-based predictions, feature selection is one of the most important steps because of redundant and non-informative features. Generally, high dimensional features contain numerous non-informative and redundant features, which affect prediction accuracy. Hence, the feature selection protocol is considered one of the most important steps in ML-based prediction (Wang et al., 2012; Manavalan et al., 2014; Manavalan and Lee, 2017; Song et al., 2017). To this end, we applied a feature selection protocol that has been proven effective in various biological applications (Manavalan and Lee, 2017; Manavalan et al., 2018), and identified the optimal features. Of those, the major contribution was from DPC (∼72%), followed by CTD, AAC, PCP, and ATC, indicating that information about the fraction of amino acids as well as their local order might play a major role in predicting PVPs. A previous study demonstrated that basic amino acids (Lys and Arg) usually occur in the flanking potential cleavage site in PVPs, as their side chain flexibility is required to accommodate the

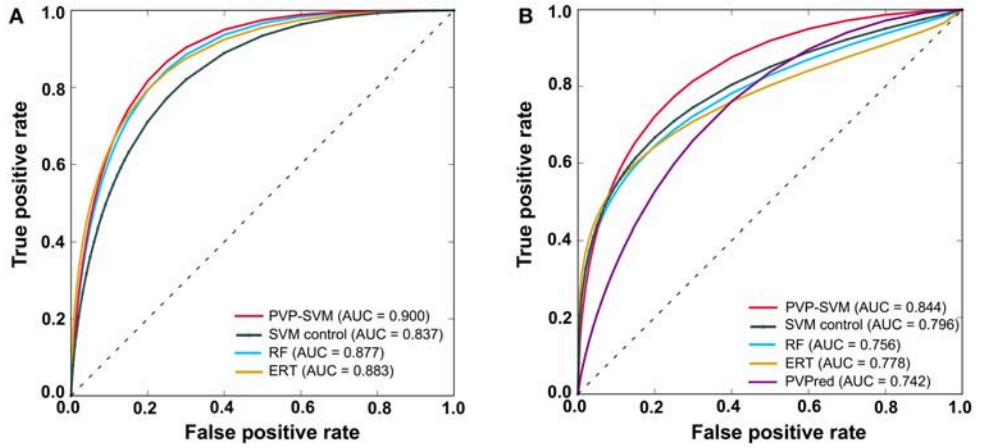


FIGURE 5 | Receiver operating characteristic curves of the prediction models. **(A)** LOOCV of the training dataset. **(B)** Evaluation with an independent dataset. Higher AUC values indicate better method performance.

change observed in the cleavage site (Coia et al., 1988; Speight et al., 1988). Interestingly, our optimal features contain these two important types of residues.

In general, if a prediction model is developed using a training dataset that contains highly homologous sequences, this method will overestimate the prediction accuracy. In this regard, Feng et al., and Ding et al., used a lower homology (<40% sequence identity) sequence dataset to develop their prediction models (Feng et al., 2013b; Ding et al., 2014). Zhang et al., developed their model using a highly homologous sequence dataset (<80% sequence identity); as a result, this method showed higher accuracy when evaluated with an independent dataset (Zhang et al., 2015). Furthermore, PVPred is the only publicly available method of the three, in the form of a web server, and was generated using the same dataset as our method. Therefore, we compared the performance of our method with PVPred only. Generally, a prediction model tends toward over-optimization in order to attain higher accuracy. Therefore, it is always necessary to evaluate the prediction model using an independent dataset, to measure the generalizability of the method (Chaudhary et al., 2016; Manavalan and Lee, 2017; Nagpal et al., 2017). Hence, we evaluated our three prediction models and PVPred on an independent dataset. Our study demonstrated that PVP-SVM consistently performed better than PVPred and the two other methods developed in this study on both datasets, indicating the greater transferability of the method.

The superior performance of PVP-SVM may be attributed to two important factors: (i) integration of previously reported features and inclusion of novel features that collectively make significant contributions to the performance; and (ii) a feature selection protocol that eliminates overlapping and redundant features. Furthermore, our approach is a general one, which is applicable to many other classification problems in structural bioinformatics. Although PVP-SVM displayed superior performance over the other methods, there is room for further improvements, including increasing the size of the training dataset based on the experimental data available in the future, incorporating novel features, and exploring different ML algorithms including stochastic gradient boosting (Xu et al., 2017) and deep learning (LeCun et al., 2015).

A user-friendly web interface has been made available, allowing researchers access to our prediction method. Indeed, this is the second method to be made publicly available, with higher accuracy than the existing method. Compared to experimental approaches, bioinformatics methods, such as PVP-SVM, represent a powerful and cost-effective approach for the proteome-wide prediction of PVPs. Therefore, PVP-SVM might be useful for large-scale PVP prediction, facilitating hypothesis-driven experimental design.

AUTHOR CONTRIBUTIONS

BM and GL conceived and designed the experiments; BM performed the experiments; BM and TS analyzed the data; BM and GL wrote paper. All authors reviewed the manuscript and agreed to this information prior to submission.

FUNDING

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education, Science, and Technology [2015R1D1A1A09060192 and 2009-0093826], and the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning [2016M3C7A1904392].

ACKNOWLEDGMENTS

The authors would like to thank Da Yeon Lee for assistance in the preparation of the manuscript.

REFERENCES

Basith, S., Manavalan, B., Gosu, V., and Choi, S. (2013). Evolutionary, structural and functional interplay of the IkappaB family members. *PLoS ONE* 8:e54178. doi: 10.1371/journal.pone.0054178

Basith, S., Manavalan, B., Govindaraj, R. G., and Choi, S. (2011). *In silico* approach to inhibition of signaling pathways of Toll-like receptors 2 and 4 by ST2L. *PLoS ONE* 6:e23989. doi: 10.1371/journal.pone.0023989

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in *Proceedings of the Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (Pittsburgh, PA: ACM).

Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinformatics* 15:120. doi: 10.1186/1471-2105-15-120

Chaudhary, K., Nagpal, G., Dhanda, S. K., and Raghava, G. P. (2016). Prediction of immunomodulatory potential of an RNA sequence for designing non-toxic siRNAs and RNA-based vaccine adjuvants. *Sci Rep*.6:20678. doi: 10.1038/srep20678

Chen, W., Feng, P. M., Lin, H., and Chou, K. C. (2014). iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed. Res. Int*. 2014:623149. doi: 10.1155/2014/623149

Chen, W., Tang, H., and Lin, H. (2017a). MethyRNA: a web server for identification of N(6)-methyladenosine sites. *J. Biomol. Struct. Dyn*. 35, 683–687. doi: 10.1080/07391102.2016.1157761

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017b). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Coia, G., Parker, M. D., Speight, G., Byrne, M. E., and Westaway, E. G. (1988). Nucleotide and complete amino acid sequences of Kunjin virus: definitive gene order and characteristics of the virus-specified proteins. *J. Gen. Virol*. 69(Pt 1), 1–21.

Deng, X., Li, J., and Cheng, J. (2013). Predicting protein model quality from sequence alignments by support vector machines. *J. Proteomics Bioinform*. S9:001. doi: 10.4172/jpb.S9-001

Ding, H., Feng, P. M., Chen, W., and Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.* 10, 2229–2235. doi: 10.1039/c4mb00316k.

Drulis-Kawa, Z., Majkowska-Skrobek, G., Maciejewska, B., Delattre, A. S., and Lavigne, R. (2012). Learning from bacteriophages - advantages and limitations of phage and phage-encoded protein applications. *Curr. Protein Pept. Sci.* 13, 699–722. doi: 10.2174/138920312804871193

Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U.S.A*. 92, 8700–8704.

Eickholt, J., Deng, X., and Cheng, J. (2011). DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. *BMC Bioinformatics* 12:43. doi: 10.1186/1471-2105-12-43

Elofsson, A., Joo, K., Keasar, C., Lee, J., Maghrabi, A. H. A., Manavalan, B., et al. (2017). Methods for estimation of model accuracy in CASP12. *Proteins* 86(Suppl. 1), 361–373. doi: 10.1101/143925

Feng, P. M., Chen, W., Lin, H., and Chou, K. C. (2013a). iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal Biochem.* 442, 118–125. doi: 10.1016/j.ab.2013.05.024

Feng, P. M., Ding, H., Chen, W., and Lin, H. (2013b). Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med*. 2013:530696. doi: 10.1155/2013/530696

Feng, P. M., Lin, H., and Chen, W. (2013c). Identification of antioxidants from sequence information using naive Bayes. *Comput. Math. Methods Med*. 2013:567529. doi: 10.1155/2013/567529

Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K. C. (2018). iDNA6mA-PseKNC: identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics*. doi: 10.1016/j.ygeno.2018.01.005. [Epub ahead of print].

Govindaraj, R. G., Manavalan, B., Basith, S., and Choi, S. (2011). Comparative analysis of species-specific ligand recognition in Toll-like receptor 8 signaling: a hypothesis. *PLoS ONE* 6:e25118. doi: 10.1371/journal.pone.0025118

Govindaraj, R. G., Manavalan, B., Lee, G., and Choi, S. (2010). Molecular modeling-based evaluation of hTLR10 and identification of potential ligands in Toll-like receptor signaling. *PLoS ONE* 5:e12713. doi: 10.1371/journal.pone.0012713

Gupta, S., Mittal, P., Madhu, M. K., and Sharma, V. K. (2017). IL17eScan: a tool for the identification of peptides inducing IL-17 response. *Front. Immunol*. 8:1430. doi: 10.3389/fimmu.2017.01430

Jara-Acevedo, R., Diez, P., Gonzalez-Gonzalez, M., Degano, R. M., Ibarrola, N., Gongora, R., et al. (2018). Screening phage-display antibody libraries using protein arrays. *Methods Mol. Biol*. 1701, 365–380. doi: 10.1007/978-1-4939-7447-4_20

Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. (2017). Assessment of model accuracy estimations in CASP12. *Proteins* 86(Suppl. 1), 345–360. doi: 10.1002/prot.25371

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., et al. (2015). An *in silico* platform for predicting, screening and designing of antihypertensive peptides. *Sci. Rep*. 5:12512. doi: 10.1038/srep12512

Lavigne, R., Ceyssens, P. J., and Robben, J. (2009). Phage proteomics: applications of mass spectrometry. *Methods Mol. Biol*. 502, 239–251. doi: 10.1007/978-1-60327-565-1_14

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* 8, 995–1005. doi: 10.1038/nrm2281

Lekunberri, I., Subirats, J., Borrego, C. M., and Balcazar, J. L. (2017). Exploring the contribution of bacteriophages to antibiotic resistance. *Environ. Pollut*. 220(Pt B), 981–984. doi: 10.1016/j.envpol.2016.11.059

Li, L., Xiong, Y., Zhang, Z.-Y., Guo, Q., Xu, Q., Liow, H.-H., et al. (2015). Improved feature-based prediction of SNPs in human cytochrome P450 enzymes. *Interdiscipl. Sci.* 7, 65–77. doi: 10.1007/s12539-014-0257-2

Lin, H., Liu, W. X., He, J., Liu, X. H., Ding, H., and Chen, W. (2015). Predicting cancerlectins by the optimal g-gap dipeptides. *Sci Rep*. 5:16964. doi: 10.1038/srep16964

Manavalan, B., Basith, S., Choi, Y. M., Lee, G., and Choi, S. (2010a). Structure-function relationship of cytoplasmic and nuclear IkappaB proteins: an *in silico* analysis. *PLoS ONE* 5:e15782. doi: 10.1371/journal.pone.0015782

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget*. 8, 77121–77136. doi: 10.18632/oncotarget.20365

Manavalan, B., Govindaraj, R., Lee, G., and Choi, S. (2011). Molecular modeling-based evaluation of dual function of IkappaBzeta ankyrin repeat domain in toll-like receptor signaling. *J. Mol. Recognit*. 24, 597–607. doi: 10.1002/jmr.1085

Manavalan, B., Kuwajima, K., Joung, I., and Lee, J. (2015). "Structure-based protein folding type classification and folding rate prediction," in *Proceedings of the Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on* (Washington, DC: IEEE).

Manavalan, B., and Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* 33, 2496–2503. doi: 10.1093/bioinformatics/btx222.

Manavalan, B., Lee, J., and Lee, J. (2014). Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS ONE* 9:e106542. doi: 10.1371/journal.pone.0106542

Manavalan, B., Murugapiran, S. K., Lee, G., and Choi, S. (2010b). Molecular modeling of the reductase domain to elucidate the reaction mechanism of reduction of peptidyl thioester into its corresponding alcohol in non-ribosomal peptide synthetases. *BMC Struct. Biol*. 10:1. doi: 10.1186/1472-6807-10-1

Manavalan, B., Shin, T. H., and Lee, G. (2018). DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 9, 1944–1956. doi: 10.18632/oncotarget.23099

Nagpal, G., Chaudhary, K., Dhanda, S. K., and Raghava, G. P. S. (2017). Computational prediction of the immunomodulatory potential of RNA sequences. *Methods Mol. Biol*. 1632, 75–90. doi: 10.1007/978-1-4939-7138-1_5

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Qiu, W. R., Sun, B. Q., Xiao, X., Xu, Z. C., and Chou, K. C. (2016). iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget* 7, 44310–44321. doi: 10.18632/oncotarget.10027.

Scholkopf, B., and Smola, A. J. (2001). *Learning with Kernels*: *Support Vector Machines*, *Regularization, Optimization*, and *Beyond*. London: MIT Press.

Seguritan, V., Alves, N. Jr., Arnoult, M., Raymond, A., Lorimer, D., Burgin, A. B. Jr., et al. (2012). Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput. Biol*. 8:e1002657. doi: 10.1371/journal.pcbi.1002657

Smola, A., and Vapnik, V. (1997). Support vector regression machines. *Adv. Neural Inf. Process. Syst*. 9, 155–161.

Song, J., Wang, H., Wang, J., Leier, A., Marquez-Lago, T., Yang, B., et al. (2017). PhosphoPredict: a bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci. Rep*. 7:6862. doi: 10.1038/s41598-017-07199-4

Speight, G., Coia, G., Parker, M. D., and Westaway, E. G. (1988). Gene mapping and positive identification of the non-structural proteins NS2A, NS2B, NS3, NS4B and NS5 of the flavivirus Kunjin and their cleavage sites. *J. Gen. Virol.* 69(Pt 1), 23–34.

Tang, H., Su, Z. D., Wei, H. H., Chen, W., and Lin, H. (2016). Prediction of cell-penetrating peptides with feature selection techniques. *Biochem. Biophys. Res. Commun*. 477, 150–154. doi: 10.1016/j.bbrc.2016.06.035

Vapnik, V. N., and Vapnik, V. (1998). *Statistical Learning Theory*. New York, NY: Wiley.

Wang, H., Feng, L., Zhang, Z., Webb, G. I., Lin, D., and Song, J. (2016). Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci. Rep*. 6:21383. doi: 10.1038/srep21383

Wang, M., Zhao, X. M., Takemoto, K., Xu, H., Li, Y., Akutsu, T., et al. (2012). FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PLoS ONE* 7:e43847. doi: 10.1371/journal.pone.0043847

Wang, Z., Tegge, A. N., and Cheng, J. (2009). Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins* 75, 638–647. doi: 10.1002/prot.22275.

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Sci.* 10(Suppl. 1), S20. doi: 10.1186/1477-5956-10-S1-S20

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H.-Y., et al. (2017). PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol*. 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Yuan, Y., and Gao, M. (2016). Proteomic analysis of a novel Bacillus jumbo phage revealing glycoside hydrolase as structural component. *Front. Microbiol*. 7:745. doi: 10.3389/fmicb.2016.00745

Zhang, D., and Tsai, J. J. P. (2005). *Machine Learning Applications in Software Engineering*. River Edge, NJ: World Scientific.

Zhang, L., Zhang, C., Gao, R., and Yang, R. (2015). An ensemble method to distinguish bacteriophage virion from non-virion proteins based on protein sequence characteristics. *Int. J. Mol. Sci.* 16, 21734–21758. doi: 10.3390/ijms160921734

Zheng, C., Wang, M., Takemoto, K., Akutsu, T., Zhang, Z., and Song, J. (2012). An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins. *PLoS ONE* 7:e49716. doi: 10.1371/journal.pone.0049716

Conflict of Interest Statement: 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.

Copyright © 2018 Manavalan, Shin and Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 14 May 2018
doi: 10.3389/fmicb.2018.00955

Identifying RNA N^6-Methyladenosine Sites in *Escherichia coli* Genome

Jidong Zhang[1], *Pengmian Feng*[2], *Hao Lin*[3]* *and Wei Chen*[3,4]*

[1] Department of Immunology, Zunyi Medical College, Zunyi, China, [2] Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, School of Public Health, North China University of Science and Technology, Tangshan, China, [3] Key Laboratory for Neuro-Information of Ministry of Education, Center for Informational Biology, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, [4] Department of Physics, Center for Genomics and Computational Biology, School of Sciences, North China University of Science and Technology, Tangshan, China

N^6-methyladenosine (m^6A) plays important roles in a branch of biological and physiological processes. Accurate identification of m^6A sites is especially helpful for understanding their biological functions. Since the wet-lab techniques are still expensive and time-consuming, it's urgent to develop computational methods to identify m^6A sites from primary RNA sequences. Although there are some computational methods for identifying m^6A sites, no methods whatsoever are available for detecting m^6A sites in microbial genomes. In this study, we developed a computational method for identifying m^6A sites in *Escherichia coli* genome. The accuracies obtained by the proposed method are >90% in both 10-fold cross-validation test and independent dataset test, indicating that the proposed method holds the high potential to become a useful tool for the identification of m^6A sites in microbial genomes.

Keywords: N^6-methyladenosine, machine learning method, nucleotide physicochemical properties, microbial genome, pseudo nucleotide composition

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Yongqiang Xing,
Inner Mongolia University of Science and Technology, China
Renzhi Cao,
Pacific Lutheran University, United States

***Correspondence:**
Hao Lin
hlin@uestc.edu.cn
Wei Chen
chenweiimu@gmail.com

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

Received: *13 March 2018*
Accepted: *24 April 2018*
Published: *14 May 2018*

Citation:
Zhang J, Feng P, Lin H and Chen W (2018) Identifying RNA N^6-Methyladenosine Sites in Escherichia coli Genome. Front. Microbiol. 9:955. doi: 10.3389/fmicb.2018.00955

INTRODUCTION

At present, ~150 kinds of RNA modifications have been found in different RNA species (Boccaletto et al., 2018), which not only enrich the genetic information, but also play critical roles in a variety of biological processes as mentioned in a recent review (Roundtree et al., 2017). Among these modifications, the N^6-methyladenosine (m^6A) is the most abundant posttranscriptional modification and has been found in the three domains of life. m^6A has been found to participate in various biological activities, such as mRNA splicing (Nilsen, 2014), mRNA translation (Wang et al., 2015), mRNA maturation (Hoernes et al., 2016), stem cell proliferation (Bertero et al., 2018), and even a series of diseases (Zhang et al., 2016; Cui et al., 2017; Li et al., 2017).

In order to reveal its biological functions, different kinds of high-throughput sequencing techniques have been proposed to map the locations of m^6A on genome wide (Dominissini et al., 2013; Linder et al., 2015; Wan et al., 2015; Hong et al., 2018). Although these techniques promoted the research progress on understanding the biological functions and the identification of RNA modifications, they are still labor-intensive and cost-ineffective. In addition, the resolution of detecting m^6A sites for most techniques is still not satisfactory. Therefore, it's necessary to develop novel methods to detect m^6A sites.

Giving the credit to the experimental data yielded by these high-throughput sequencing techniques as reported in a recent work (Chen X. et al., 2017), some machine learning based

computational methods have been proposed to identify m^{6}A sites (Chen et al., 2015a,b, 2016a, 2017b,c; Zhou et al., 2016). Although these methods are really good complements to experimental methods for detecting m^{6}A sites, to the best of our knowledge, so far there is no computational tool available for detecting m^{6}A sites in microbial genomes.

Stimulated by the successful applications of machine learning methods in computational genomics and proteomics (Chen et al., 2012; Feng et al., 2013; Cao et al., 2016, 2017a,b; Hu et al., 2018), in the present work, we presented a support vector machine (SVM) based method for identifying m^{6}A sites in the *Escherichia coli* (*E. coli*) genome. By encoding the RNA sequences using nucleotide chemical property and accumulated nucleotide frequency, the proposed method obtained promising performances in 10-fold cross validation test. Moreover, we also validated the method on the independent dataset and obtained satisfactory results.

MATERIALS AND METHODS

Benchmark Dataset

The m^{6}A site containing sequences of *E. coli* genome were obtained from the RMBase database (Xuan et al., 2018). All the sequences are 41 bp long with the m^{6}A site in the center. To overcome redundancy and reduce the homology bias, sequences with more than 80% sequence similarity were removed by using the CD-HIT program (Fu et al., 2012). After such a screening procedure, 2,055 m^{6}A site containing sequences were retained and regarded as positive samples.

The negative samples (non-m^{6}A site containing sequences) were obtained by choosing the 41-bp long sequences with the central adenosine that was not experimentally confirmed occurring methylation on its 6th nitrogen. By doing so, we could obtain a large number of negative samples. After removing sequences with identify >80%, the number of negative samples are still dramatically larger than that of positive samples. To balance out the numbers between positive and negative samples in model training, we randomly picked out the same number of negative samples and repeated this process 10 times. Therefore, 10 negative subsets were obtained, and each of them includes 2,055 non-m^{6}A site containing sequences. The positive and negative samples thus obtained are provided in Supplementary Material.

Sequence Encoding Scheme

Inspired by recent studies (Chen et al., 2016b,c,d, 2017a,d; Feng et al., 2017), in order to transfer the RNA sequences into discrete vectors that can be recognized and handled by machine learning methods, we encoded RNA sequences using nucleotide chemical properties and accumulated nucleotide frequency. Their brief descriptions are as following.

The four nucleotides, namely, adenine (A), guanine (G), cytosine (C), and uracil (U) can be classified into three different groups according to their physicochemical properties, i.e., ring structures, secondary structures, and chemical functionality (Chen et al., 2016b,c,d, 2017a,d; Feng et al., 2017). Therefore, based on the different physicochemical properties, the four coordinates (1, 1, 1), (0, 0, 1), (1, 0, 0), and (0, 1, 0) were used to represent the four bases (A, C, G, and U) of RNA, respectively.

In order to include nucleotide composition surrounding the modification site as well, the accumulated nucleotide frequency of any nucleotide n_j at position i was also used to represent RNA sequences and was defined as

$$d_i = \frac{1}{|N_i|}\sum_{j=1}^{l} f(n_j),\ f\left(n_j\right) = \begin{cases} 1 \ if\ n_j = q \\ 0 \ other\ cases \end{cases} \quad (1)$$

where $|N_i|$ is the length of the sliding substring concerned, l denotes each of the site locations counted in the substring, $q\epsilon${A, C, G, U}.

By integrating both nucleotide physicochemical properties and accumulated nucleotide frequency, an L *nt* long RNA sequence could be represented a $4L$-dimensional vector (Chen et al., 2016b,c,d, 2017a,d; Feng et al., 2017).

Support Vector Machine

As an efficient supervised machine learning algorithm, SVM has been widely used in the realm of bioinformatics (Cao et al., 2014; Li et al., 2017; Wang et al., 2017b; Zhang et al., 2017). Its basic idea is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane.

In the current study, the implementation of SVM was performed by using the LibSVM package 3.18, available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The radial basis kernel function (RBF) was used to obtain the classification hyperplane. The grid search method was applied to optimize its regularization parameter C and kernel parameter γ.

Evaluation Metrics

The performance was evaluated by using the following four metrics, namely sensitivity (*Sn*), specificity (*Sp*), Accuracy (*Acc*), and the Mathew's correlation coefficient (*MCC*), which can be expressed as

$$\begin{cases} Sn = \frac{TP}{TP+FN} \times 100\% \\ Sp = \frac{TN}{TN+FP} \times 100\% \\ Acc = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \\ MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FN)\times(TN+FP)}} \end{cases} \quad (2)$$

where *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively.

To further evaluate the performance of the current method more objectively, inspired by recent works (Wang et al., 2017a), the ROC (receiver operating characteristic) curve was also plotted. Its vertical coordinate indicates the true positive rate (sensitivity) and the horizontal coordinate indicates the false positive rate (1-specificity). The area under the ROC curve (auROC) is an indicator of the performance quality of a binary classifier, i.e., the value 0.5 of auROC is equivalent to random prediction while the value 1 of auROC represents a perfect one.

TABLE 1 | The 10-fold cross validation predictive results by using different negative datasets for identifying m^6A sites in *E. coli*.

Dataset	Sn (%)	Sp (%)	Acc (%)	MCC
Negative set 1	100.00	98.59	99.29	0.98
Negative set 2	100.00	98.78	99.39	0.98
Negative set 3	100.00	98.44	99.22	0.98
Negative set 4	100.00	98.88	99.44	0.98
Negative set 5	100.00	98.44	99.22	0.98
Negative set 6	100.00	98.49	99.25	0.98
Negative set 7	100.00	98.54	99.27	0.98
Negative set 8	100.00	98.69	99.34	0.98
Negative set 9	100.00	98.49	99.25	0.98
Negative set 10	100.00	98.25	99.12	0.97
Average	100.00	98.56	99.28	0.98

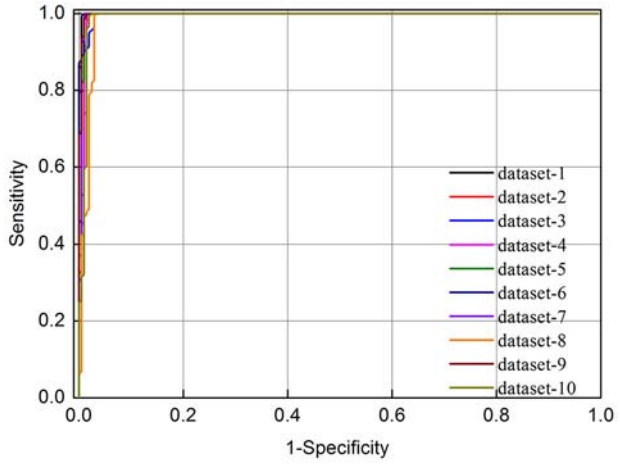


FIGURE 1 | The ROC curves of 10-fold cross validation test for identifying m^6A sites in *E. coli* based on different negative datasets. The vertical coordinate is the true positive rate (*Sn*) while horizontal coordinate is the false positive rate (1-*Sp*).

RESULTS AND DISCUSSIONS

Performance for m^6A Site Identification

In statistical prediction, independent dataset test, *K*-fold cross-validation test and jackknife test are often used to derive the metric values for a predictor (Chou, 2011). In order to saving computational time, the 10-fold cross-validation test was used to examine the performance of the proposed method. In 10-fold cross-validation test, the samples in the dataset are randomly partitioned into 10 equal sized sub-datasets. Of the 10 sub-datasets, a single sub-dataset is retained as the validation data for testing the model, and the remaining 9 sub-datasets are used as training data. The process is then repeated 10 times, with each of the 10 sub-datasets used exactly once as the validation data.

By encoding RNA sequences using nucleotide chemical property and accumulated nucleotide frequency, each sample in the dataset was represented by a $(4 \times 41) = 164$-dimensional vector and used as the input of SVM. The 10-fold cross-validation test results for identifying m^6A sites in *E. coli* were listed in **Table 1**. In addition, to demonstrate that whether its accuracy is sensitive to the selection of negative data, the method was also tested on the other nine negative datasets, respectively. Their predictive results of the 10-fold cross-validation were also provided in **Table 1**.

As indicated in **Table 1**, we found that the predictive accuracy is not affected by the selection of negative data. In addition, the 10 ROC curves obtained based on the 10 different negative datasets were also plotted in **Figure 1**. It was found that their auROCs are all higher than 0.98. These results demonstrate the reliability and robustness of the model developed in this study.

Comparison With Other Methods

In order to demonstrate the effectiveness of nucleotide chemical property and accumulated nucleotide frequency for identifying m^6A sites in *E. coli*, we compared the performance of the proposed method with that of the method based on other commonly used RNA sequence features. Chen et al. have proposed the pseudo nucleotide composition (PseKNC) to represent RNA sequences (Chen et al., 2014a,b), in which both the local and global sequence order information w included. Since it has been proposed in 2014, PseKNC have been used in in many branches of computational genomics (Guo et al., 2014; Lin et al., 2014, 2017). Therefore, we employed the SVM to perform the comparisons between the model based on nucleotide chemical property and accumulated nucleotide frequency features and that based on the PseKNC features (Chen et al., 2015a). The 10-fold cross-validation test results were listed in **Table 2**.

TABLE 2 | Comparison of different parameters for identifying m^6A sites in *E. coli*.

Parameters	Sn (%)	Sp (%)	Acc (%)	MCC
PseKNC	65.74	60.29	63.02	0.26
Secondary structure	67.06	60.73	63.89	0.28
Our method	100.00	98.56	99.28	0.98

As indicated in a recent study (Schwartz et al., 2013), the m^6A modification is also affected by RNA secondary structures. Therefore, we performed the prediction of m^6A sites by using RNA secondary structure. To this end, all the sequences in the benchmark dataset were encoded by using their second structures. The details about the encoding scheme based on secondary structures can be found in a recent work (Xue et al., 2005). By doing so, each RNA sequence is converted to a 32 dimensional vector (Xue et al., 2005) and used as the input feature of SVM. Its 10-fold cross-validation test results were also listed in **Table 2**.

As shown in **Table 2**, the predictive performance of the method based on nucleotide chemical property and accumulated nucleotide frequency is dramatically higher than that based on PseKNC and RNA secondary structure.

Validation on Independent Dataset

The proposed method trained based on the benchmark dataset from the *E. coli* genome was further used to identify the m^6A sites in the *P. aeruginosa* genome. For this purpose, we firstly collected the 5,814 experimentally confirmed m^6A sites from the RMBase to form an independent dataset, which is given in Supporting Information S2. Of the 5,814 m^6A sites in the *P. aeruginosa*, 5,809 were correctly identified, indicating that the proposed method is really quite promising for identifying m^6A sites in microbial genomes.

CONCLUSION

In this study, we present a computational method to identify m^6A sites in the *E. coli* genome by encoding the RNA sequences using nucleotide chemical property and accumulated nucleotide frequency. The results obtained based on the benchmark dataset and independent dataset demonstrate that the proposed method is powerful and promising in discovering m^6A sites. We hope that the proposed method will be helpful for the future research on m^6A sites in microbial genomes.

Since user-friendly and publicly accessible web-servers (Feng et al., 2018)and databases (Liang et al., 2017) represent the direction of developing new prediction method, we will make efforts in our future work to provide a web-server for the method presented in this paper.

AUTHOR CONTRIBUTIONS

HL and WC: conceived and designed the experiments; JZ and PF: performed the experiments; HL and WC: wrote the paper.

ACKNOWLEDGMENTS

This work was supported by the National Nature Science Foundation of China (Nos. 31771471, 61772119), Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028), the Outstanding Youth Foundation of North China University of Science and Technology (No. JP201502), and the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015Z006, ZYGX2016J125, ZYGX2016J118), Natural Science Foundation of Guizhou Province (QKH-2016-1167); The Scientific and Technological Innovation Project for Oversea Students of Guizhou province (QR-2016-20); High School Science and Technology Talent Support Project of Guizhou Province (QJH-KY-2016-079).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00955/full#supplementary-material

REFERENCES

Bertero, A., Brown, S., Madrigal, P., Osnato, A., Ortmann, D., Yiangou, L., et al. (2018). The SMAD2/3 interactome reveals that TGFβ controls m(6)A mRNA methylation in pluripotency. *Nature* 555, 256–259. doi: 10.1038/nature25784

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Cao, R., Adhikari, B., Bhattacharya, D., Sun, M., Hou, J., and Cheng, J. (2017a). QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics* 33, 586–588. doi: 10.1093/bioinformatics/btw694

Cao, R., Bhattacharya, D., Hou, J., and Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics* 17:495. doi: 10.1186/s12859-016-1405-y

Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., and Chen, Z. (2017b). ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 22:E1732. doi: 10.3390/molecules22101732

Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinformatics* 15:120. doi: 10.1186/1471-2105-15-120

Chen, W., Feng, P., Ding, H., and Lin, H. (2016a). Identifying N (6)-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. *Mol. Genet. Genomics* 291, 2225–2229. doi: 10.1007/s00438-016-1243-7

Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.-C. (2015a). iRNA-methyl: identifying N-6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* 490, 26–33. doi: 10.1016/j.ab.2015.08.021

Chen, W., Feng, P., Tang, H., Ding, H., and Lin, H. (2016b). Identifying 2′-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. *Genomics* 107, 255–258. doi: 10.1016/j.ygeno.2016.05.003

Chen, W., Feng, P., Tang, H., Ding, H., and Lin, H. (2016c). RAMPred: identifying the N-1-methyladenosine sites in eukaryotic transcriptomes. *Sci. Rep.* 6:31080. doi: 10.1038/srep31080

Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.-C. (2017a). iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget* 8, 4208–4217. doi: 10.18632/oncotarget.13758

Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., and Chou, K.-C. (2014a). PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* 456, 53–60. doi: 10.1016/j.ab.2014.04.001

Chen, W., Lin, H., Feng, P. M., Ding, C., Zuo, Y. C., and Chou, K. C. (2012). iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS ONE* 7:e47843. doi: 10.1371/journal.pone.0047843

Chen, W., Tang, H., and Lin, H. (2017b). MethyRNA: a web server for identification of N-6-methyladenosine sites. *J. Biomol. Struct. Dyn.* 35, 683–687. doi: 10.1080/07391102.2016.1157761

Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K.-C. (2016d). iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* 5, 155–163. doi: 10.1038/mtna.2016.37

Chen, W., Tran, H., Liang, Z., Lin, H., and Zhang, L. (2015b). Identification and analysis of the N-6-methyladenosine in the Saccharomyces cerevisiae transcriptome. *Sci. Rep.* 5: 13859 .doi: 10.1038/srep13859

Chen, W., Xing, P., and Zou, Q. (2017c). Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. *Sci. Rep.* 7:40242. doi: 10.1038/srep40242

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017d). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.-C. (2014b). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* 31, 119–120. doi: 10.1093/bioinformatics/btu602

Chen, X., Sun, Y. Z., Liu, H., Zhang, L., Li, J. Q., and Meng, J. (2017). RNA methylation and diseases: experimental results, databases, web servers and computational models. *Brief Bioinform*. doi: 10.1093/bib/bbx142. [Epub ahead of print].

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273, 236–247. doi: 10.1016/j.jtbi.2010.12.024

Cui, Q., Shi, H., Ye, P., Li, L., Qu, Q., Sun, G., et al. (2017). m(6)A RNA methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. *Cell Rep*. 18, 2622–2634. doi: 10.1016/j.celrep.2017.02.059

Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N., and Rechavi, G. (2013). Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing. *Nat. Protoc.* 8, 176–189. doi: 10.1038/nprot.2012.148

Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., and Chou, K.-C. (2017). iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol. Ther. Nucleic Acids* 7, 155–163. doi: 10.1016/j.omtn.2017.03.006

Feng, P. M., Chen, W., Lin, H., and Chou, K. C. (2013). iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* 442, 118–125. doi: 10.1016/j.ab.2013.05.024

Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K. C. (2018). iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* doi: 10.1016/j.ygeno.2018.01.005. [Epub ahead of print].

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

Guo, S.-H., Deng, E.-Z., Xu, L.-Q., Ding, H., Lin, H., Chen, W., et al. (2014). iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* 30, 1522–1529. doi: 10.1093/bioinformatics/btu083

Hoernes, T. P., Huttenhofer, A., and Erlacher, M. D. (2016). mRNA modifications: dynamic regulators of gene expression? *RNA Biol.* 13, 760–765. doi: 10.1080/15476286.2016.1203504

Hong, T., Yuan, Y., Chen, Z., Xi, K., Wang, T., Xie, Y., et al. (2018). Precise antibody-independent m6A identification via 4SedTTP-involved and FTO-assisted strategy at single-nucleotide resolution. *J. Am. Chem. Soc.* doi: 10.1021/jacs.7b13633. [Epub ahead of print].

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol*. doi: 10.1080/15476286.2018.1457935. [Epub ahead of print].

Li, Z., Weng, H., Su, R., Weng, X., Zuo, Z., Li, C., et al. (2017). FTO plays an oncogenic role in acute myeloid leukemia as a N(6)-methyladenosine RNA demethylase. *Cancer Cell* 31, 127–141. doi: 10.1016/j.ccell.2016.11.017

Liang, Z. Y., Lai, H. Y., Yang, H., Zhang, C. J., Yang, H., Wei, H. H., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics* 33, 467–469. doi: 10.1093/bioinformatics/btw630

Lin, H., Deng, E.-Z., Ding, H., Chen, W., and Chou, K.-C. (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* 42, 12961–12972. doi: 10.1093/nar/gku1019

Lin, H., Liang, Z.-Y., Tang, H., and Chen, W. (2017). Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2017.2666141. [Epub ahead of print].

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., and Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* 12, 767–772. doi: 10.1038/nmeth.3453

Nilsen, T. W. (2014). Molecular biology. Internal mRNA methylation finally finds functions. *Science* 343, 1207–1208. doi: 10.1126/science.1249340

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

Schwartz, S., Agarwala, S. D., Mumbach, M. R., Jovanovic, M., Mertins, P., Shishkin, A., et al. (2013). High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* 155, 1409–1421. doi: 10.1016/j.cell.2013.10.047

Wan, Y., Tang, K., Zhang, D., Xie, S., Zhu, X., Wang, Z., et al. (2015). Transcriptome-wide high-throughput deep m(6)A-seq reveals unique differential m(6)A methylation patterns between three organs in *Arabidopsis thaliana*. *Genome Biol.* 16:272. doi: 10.1186/s13059-015-0839-2

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., et al. (2017a). LRLSHMDA: Laplacian Regularized Least Squares for Human Microbe-Disease Association prediction. *Sci. Rep.* 7:7601. doi: 10.1038/s41598-017-08127-2

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., et al. (2015). N(6)-methyladenosine modulates messenger RNA translation efficiency. *Cell* 161, 1388–1399. doi: 10.1016/j.cell.2015.05.014

Wang, Y., You, Z., Li, X., Chen, X., Jiang, T., and Zhang, J. (2017b). PCVMZM: using the probabilistic classification vector machines model combined with a zernike moments descriptor to predict protein-protein interactions from protein sequences. *Int. J. Mol. Sci.* 18:1029. doi: 10.3390/ijms18051029

Xuan, J. J., Sun, W. J., Lin, P. H., Zhou, K. R., Liu, S., Zheng, L. L., et al. (2018). RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res.* 46, D327–D334. doi: 10.1093/nar/gkx934.

Xue, C., Li, F., He, T., Liu, G. P., Li, Y., and Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 6:310. doi: 10.1186/1471-2105-6-310

Zhang, C., Samanta, D., Lu, H., Bullen, J. W., Zhang, H., Chen, I., et al. (2016). Hypoxia induces the breast cancer stem cell phenotype by HIF-dependent and ALKBH5-mediated m(6)A-demethylation of NANOG mRNA. *Proc. Natl. Acad. Sci. U.S.A.* 113, E2047–E2056. doi: 10.1073/pnas.1602883113

Zhang, L., Ai, H., Chen, W., Yin, Z., Hu, H., Zhu, J., et al. (2017). CarcinoPred-EL: novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. *Sci. Rep.* 7:2118. doi: 10.1038/s41598-017-02365-0

Zhou, Y., Zeng, P., Li, Y. H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.* 44:e91. doi: 10.1093/nar/gkw104

Conflict of Interest Statement: 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.

Copyright © 2018 Zhang, Feng, Lin and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 20 June 2018
doi: 10.3389/fmicb.2018.01338

Faba Bean (*Vicia faba* L.) Nodulating Rhizobia in Panxi, China, Are Diverse at Species, Plant Growth Promoting Ability, and Symbiosis Related Gene Levels

Yuan X. Chen[1†], *Lan Zou*[1†], *Petri Penttinen*[2,3†], *Qiang Chen*[1], *Qi Q. Li*[1], *Chang Q. Wang*[1] *and Kai W. Xu*[1*]

[1] College of Resources, Sichuan Agricultural University, Chengdu, China, [2] Zhejiang Provincial Key Laboratory of Carbon Cycling in Forest Ecosystems and Carbon Sequestration, School of Environmental & Resource Sciences, Zhejiang Agriculture & Forestry University, Lin'an, China, [3] Ecosystems and Environment Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland

OPEN ACCESS

Edited by:
Xing Chen,
China University of Mining and Technology, China

Reviewed by:
Margarita Kambourova,
Institute of Microbiology (BAS), Bulgaria
Prasun Ray,
Noble Research Institute, LLC, United States

***Correspondence:**
Kai W. Xu
xkwei@sicau.edu.cn

[†]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: 17 February 2018
Accepted: 31 May 2018
Published: 20 June 2018

Citation:
Chen YX, Zou L, Penttinen P, Chen Q, Li QQ, Wang CQ and Xu KW (2018) Faba Bean (Vicia faba L.) Nodulating Rhizobia in Panxi, China, Are Diverse at Species, Plant Growth Promoting Ability, and Symbiosis Related Gene Levels. Front. Microbiol. 9:1338. doi: 10.3389/fmicb.2018.01338

We isolated 65 rhizobial strains from faba bean (*Vicia faba* L.) from Panxi, China, studied their plant growth promoting ability with nitrogen free hydroponics, genetic diversity with clustered analysis of combined ARDRA and IGS-RFLP, and phylogeny by sequence analyses of 16S rRNA gene, three housekeeping genes and symbiosis related genes. Eleven strains improved the plant shoot dry mass significantly comparing to that of not inoculated plants. According to the clustered analysis of combined ARDRA and IGS-RFLP the isolates were genetically diverse. Forty-one of 65 isolates represented *Rhizobium anhuiense*, and the others belonged to *R. fabae*, *Rhizobium vallis*, *Rhizobium sophorae, Agrobacterium radiobacter,* and four species related to *Rhizobium* and *Agrobacterium*. The isolates carried four and five genotypes of *nifH* and *nodC*, respectively, in six different *nifH-nodC* combinations. When looking at the species-*nifH-nodC* combinations it is noteworthy that all but two of the six *R. anhuiense* isolates were different. Our results suggested that faba bean rhizobia in Panxi are diverse at species, plant growth promoting ability and symbiosis related gene levels.

Keywords: faba bean, rhizobia, genetic diversity, multilocus sequence analysis, symbiosis gene, lateral gene transfer

INTRODUCTION

Legumes like faba bean (*Vicia faba* L.) and rhizobial bacteria can form a symbiotic relationship in which the legume host provides the rhizobia with nutrients and niche while rhizobia provide the host with fixed atmospheric dinitrogen in the form of ammonia. Owing to symbiosis, legumes can act as pioneer plants in nitrogen deficient areas and improve soil fertility (Graham and Vance, 2003; Gentzbittel et al., 2015). Nitrogen fertilization affects the environment; however, applying biological N fixation (BNF) has some advantages over synthetic N fertilizers. If incorporated into the soil, legumes do not acidify the soil like ammonium-based fertilizers (Crews and Peoples, 2004). Unlike the production of synthetic N fertilizers, BNF does not rely on non-renewable energy sources (Crews and Peoples, 2004).

In legumes, nitrogen fixation takes place in a specific root or stem organ called nodule. The formation of plant growth promoting symbiosis requires that the legume and the rhizobia are compatible, and that the rhizobia fix nitrogen efficiently. Inoculating the legume with suitable rhizobia increases growth when compatible rhizobia are not present or when the compatible rhizobia are not efficient (Thilakarathna and Raizada, 2017).

Faba bean, a grain legume grown worldwide, is a good resource of protein, starch, cellulose and minerals. Its high yield and great adaption to different environments makes faba bean very popular among farmers, feed and food manufacturers (Haciseferoğullar et al., 2003). Moreover, the capacity for biological nitrogen fixation with rhizobial bacteria makes faba bean a renewable resource for sustainable agriculture (Köpke and Nemecek, 2010). Thus, it is common that faba bean is grown as an intercrop or in rotation with non-legume plants (Song et al., 2007; Mei et al., 2012). However, in China faba bean frequently receive synthetic N fertilizer, resulting in over fertilization (Li et al., 2016).

Panxi region in Sichuan, southwestern China, is on the western margin of Yangtze Block, between Tibet Plateau, Yunnan-Guizhou Plateau and Sichuan basin. Panxi is within the South-West China mountains biodiversity hotspot (Wu et al., 2006; www.cepf.net/resources/hotspots/Asia-Pacific/Pages/Mountains-of-Southwest-China.aspx). Mountains occupy 80% of the total area of Panxi and the altitude differences in this area reach 5,600 m. Panxi receives plenty of rainfall and strong solar radiation, and the climate ranges from southern Asian semitropical climate to northern temperate climate with xerothermic climate as the main characteristic of the arid-hot river valley area. Faba bean is one of the main crops in Panxi. Cultivation relies on seeds produced by farmers themselves. N fertilizers would be unnecessary if the soils hosted compatible, plant growth promoting rhizobia.

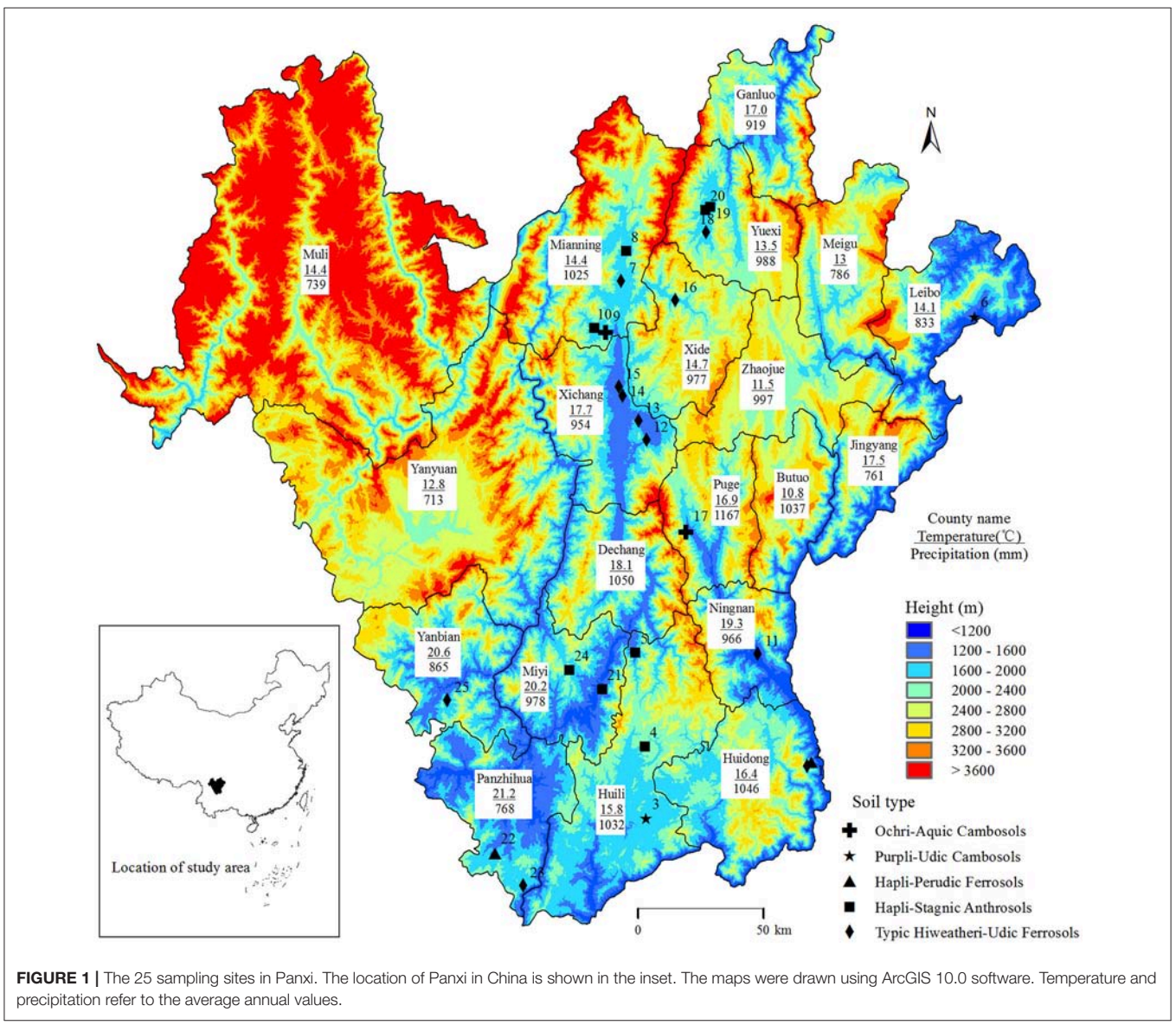


FIGURE 1 | The 25 sampling sites in Panxi. The location of Panxi in China is shown in the inset. The maps were drawn using ArcGIS 10.0 software. Temperature and precipitation refer to the average annual values.

The species range of rhizobia nodulating legumes in Panxi differs from that in other parts of China. For example, in Panxi *Leucaena leucocaphala* and *Pueraria lobate* were mostly nodulated by *Ensifer* and *Rhizobium* strains, respectively, while in subtropical China *L. leucocaphala* was nodulated by *Mesorhizobium* strains and in other parts of Sichuan *P. lobate* by *Bradyrhizobium* strains (Chen et al., 2004; Wang et al., 2006; Xie et al., 2009; Xu et al., 2013). Since faba bean rhizobia in Panxi have not been studied systematically prior to this study, our aim was to assess if faba bean rhizobia in the area were diverse and unique. Thus, we isolated rhizobia from faba bean growing in the special arid-hot environment of Panxi in diverse soil types, and studied their plant growth promoting ability, genetic diversity and phylogeny based on molecular methods.

MATERIALS AND METHODS

Isolation of Strains

Local variety faba bean samples were taken in 25 sites in Panxi, Sichuan, China (**Figure 1**) to collect root nodules. Nodules were surface sterilized in 95% ethanol for 3 min and 0.1% $HgCl_2$ for 5 min, followed by rinsing six times in sterile distilled water. The sterilized nodules were crushed individually and streaked on yeast extract mannitol (YEM) medium (Vincent, 1970) containing 25 mg L^{-1} congo red at 28°C. The purified strains were stored on YEM slants at 4°C for short term and in 25% glycerol at −80°C for long term storage.

Nodulation Assays

The nodulation ability and symbiotic efficiency of the isolates was tested on the local faba bean (*V. faba* L.) cultivar Hanyuan dabaidou. Seeds of faba bean were immersed in 95% ethanol for 5 min, rinsed for 5 min with 0.2% mercury bichloride ($HgCl_2$) and 8 times (10 min per time) with sterilized water. After surface sterilization, the seeds were soaked in sterilized water overnight to soften the thick and hard seed coat. The seeds were transferred on 0.5% water-agar for germination. The seedlings were transplanted in sterile 250 ml infusion bottles containing Jensen's solution (Vincent, 1970) in all inoculation assays. The seedlings were inoculated with 1.5 ml of the culture containing ca 10^9 bacterial cells per milliliter and grown under a 16 h light and 8 h dark regime at 25°C in greenhouse. The assays were done in triplicate with one seedling per bottle, including the uninoculated controls. After 50 days, the plants were harvested and the numbers of nodules and the plant shoot dry mass were measured. One-way analysis of variance with a least significant difference (LSD) analysis ($P = 0.05$) was done using Excel 2010 (Microsoft, Redmond, USA) and SPSS 17.0 (SPSS Inc., Chicago, USA).

PCR-RFLP and CACAI

Total DNA was extracted by GUTC (Guanidinium-Tris-CDTA buffer with celite) method (Terefework et al., 2001) from purified bacteria. 16S rDNA and intergenic spacer region (IGS) of the strains were amplified for restriction fragment

TABLE 1 | PCR primers and reaction procedures applied in this study.

Gene	Primers	Reaction procedure	References
16S rDNA	P1:5′-AGAGTTTGATCCTGGCTCAGAACGAACGCT-3′; P6: 5′- TACGGCTACCTTGTTACGACTTCACCCC-3′	92°C for 3 min, 30 cycles of 94°C for 1 min, 58°C for 1 min, 72°C for 2 min, final extension for 72°C 8 min	Tan et al., 1997
IGS	pHr (F):5′-TGCGGCTGGATCACCTCCTT-3′; p23SR01(R):5′-GGCTGC TTCTAAGCCAAC-3′	92°C for 3 min, 30 cycles of 94°C for 1 min, 60°C for 1 min, 72°C for 2 min, final extension for 72°C 8 min	Navarro et al., 1992; Massol-Deya et al., 1995
atpD	atpD255F: 5′-GCTSGGCCGCATCMTSAACGTC-3′; atpD782R: 5′-GCCGACACTTCMGAACCNGCCTG-3′	95°C for 2 min; 30 cycles of 94°C for 45 s, 59°C for 1 min, 72°C for 1.5 min; final extension 72°C for 10 min	Vinuesa et al., 2005
glnII	glnII12F:YAAGCTCGAGTACATYTGGCT; glnII689R: TGCATGCCSGAGCCGTTCCA	95°C for 5 min; 30 cycles of 94°C for 1 min, 58°C for 1 min, 72°C for 1 min; final extension 72°C for 10 min	Vinuesa et al., 2005
recA	recA 41F: 5′-TTCGGCAAGGGMTCGRTSATG-3′; recA 640R: 5′-ACATSACRCCGATCTTCATGC-3′	95°C for 5 min; 30 cycles of 94°C for 1 min, 58°C for 1 min, 72°C for 1 min; final extension 72°C for 10 min	Vinuesa et al., 2005
nifH	nifHctg:5′-CTCATCGTCGGCTGTGACCC-3′; nifHI: 5′-AGCATGTCYTCSAGYTCNTCCA-3′	95°C for 3 min; 40 cycles of 94°C for 1 min, 59°C for 1 min, 72°C for 1 min; final extension 72°C for 5 min	Laguerre et al., 2001; Gurkanli et al., 2014
	nifH1F:5′-GTCTCCTATGACGTGCTCGG-3′; nifH1R:5′-GCTTCCATGGTGATCGGGGT-3′	94°C for 3 min; 30 cycles of 94°C for 30 s, 58°C for 30 s, 72°C for 1 min; final extension 72°C for 5 min	Rivas et al., 2002
nodC	nodC540: 5′-TGATYGAYATGGARTAYTGGYT-3′; nodC1160: 5′-CGYGACAGCCANTCKCTATTG-3′	95°C for 5 min; 30 cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 1 min; final extension 72°C for 10 min	Sarita et al., 2005
	nofCf: 5′-GCTGCCTATGCAGACGATG-3′; nodCr: 5′-GGTTACTGGCTTTCATTTGGC-3′	94°C for 5 min; 30 cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 3 min; final extension 72°C for 7 min	Moschetti et al., 2005

TABLE 2 | Rhizobial isolates from faba bean in Panxi, their genetic and symbiotic characteristics and phylogenetic affiliation.

Isolate[a]	Sampling site[b]	16S rDNA RFLP genotype[c]	IGS RFLP genotype[c]	CACAI genotype	CACAI group[d]	MLSA[e]	Plant shoot dry mass (g plant^{-1})[f]	No. of nodules per plant
SCAUf129	21	a	1	1	A		1.593 ± 0.340↑*	34.3
SCAUf141	17	a	17	5	A		1.523 ± 0.020↑*	41.7
SCAUf142	4	a	1	1	A		1.457 ± 0.247↑*	93.0
SCAUf110	11	a	21	9	B		1.397 ± 0.301↑*	18.7
SCAUf126	12	a	1	1	A		1.377 ± 0.074↑*	25.0
SCAUf118	7	a	1	1	A		1.375 ± 0.025↑*	62.0
SCAUf114	12	a	1	1	A		1.327 ± 0.186↑*	40.3
SCAUf122	14	a	1	1	A		1.287 ± 0.156↑*	39.7
SCAUf123	9	g	22	24	C		1.283 ± 0.063↑*	43.3
SCAUf113	22	a	1	1	A		1.280 ± 0.265↑*	94.3
SCAUf148	16	h	20	25	D		1.267 ± 0.201↑*	52.0
SCAUf131	25	a	1	1	A	*R. anhuiense* CCBAU 23252^T (98.5%)	1.215 ± 0.365	98.5
SCAUf125	18	a	23	10	E		1.135 ± 0.095	61.5
SCAUf136	9	a	1	1	A		1.130 ± 0.160	58.7
SCAUf111	21	a	1	1	A		1.127 ± 0.109	72.3
SCAUf133	11	a	19	7	B	*R. sophorae* CCBAU 03386^T (96.6 %)	1.120 ± 0.210	17.0
SCAUf102	2	a	1	1	A		1.097 ± 0.138	28.3
SCAUf106	1	a	5	13	F	*R. vallis* CCBAU 65647^T (93.0%)	1.095 ± 0.088	48.0
SCAUf109	1	a	4	12	G	*R. vallis* CCBAU 65647^T (93.0%)	1.094 ± 0.087	46.7
SCAUf139	4	a	1	1	A		1.087 ± 0.121	40.3
SCAUf150	16	h	20	25	D	*A. radiobacter* NCPPB 2437^T (97.3%)	1.080 ± 0.131	4.0
SCAUf121	23	a	18	6	A		1.050 ± 0.044	44.7
SCAUf134	15	a	19	7	B		1.017 ± 0.153	27.0
SCAUf115	7	a	1	1	A		0.950 ± 0.410	22.5
SCAUf87	19	c	8	19	H	*A. radiobacter* NCPPB 2437^T (97.3%)	0.934 ± 0.146	23.0
SCAUf137	23	a	18	6	A		0.933 ± 0.206	35.0
SCAUf101	2	a	1	1	A		0.930 ± 0.021	27.7
SCAUf143	15	a	19	7	B		0.923 ± 0.228	22.5
SCAUf147	15	a	19	7	B		0.893 ± 0.173	22.0
SCAUf146	12	a	1	1	A		0.870 ± 0.150	4.5
SCAUf99	20	a	6	14	E	*R. sophorae* CCBAU 03386^T (96.6 %)	0.864 ± 0.041	59.7
SCAUf92	1	a	2	8	A		0.856 ± 0.099	23.7
SCAUf132	21	a	1	1	A		0.833 ± 0.231	16.7
SCAUf103	19	f	15	22	I		0.820 ± 0.066	68.0
SCAUf107	13	a	12	4	J		0.815 ± 0.066	3.0
SCAUf149	11	i	25	26	K	*A. radiobacter* NCPPB 2437^T (97.3%)	0.805 ± 0.155	35.0
CK		–	–	–			0.801±0.139	0.0
SCAUf95	5	a	1	1	A		0.790 ± 0.169	20.0
SCAUf120	17	a	17	5	A		0.790 ± 0.030	25.0
SCAUf90	10	a	11	3	B	*R. sophorae* CCBAU 03386^T (96.6 %)	0.785 ± 0.051	60.3
SCAUf94	8	a	9	16	L	*R. vallis* CCBAU 65647^T (93.0%)	0.780 ± 0.060	18.7

(Continued)

TABLE 2 | Continued

Isolate[a]	Sampling site[b]	16S rDNA RFLP genotype[c]	IGS RFLP genotype[c]	CACAI genotype	CACAI group[d]	MLSA[e]	Plant shoot dry mass (g plant^{-1})[f]	No. of nodules per plant
SCAUf86	13	b	12	17	J	*R. sophorae* CCBAU 03386^T (99.3 %)	0.770 ± 0.121	12.7
SCAUf124	14	a	1	1	A		0.753 ± 0.067	36.0
SCAUf138	14	a	1	1	A		0.750 ± 0.057	38.0
SCAUf145	18	a	17	5	A		0.750 ± 0.110	35.5
SCAUf140	17	a	18	6	A	*R. anhuiense* CCBAU 23252^T (98.9%)	0.743 ± 0.170	23.0
SCAUf117	22	a	1	1	A		0.737 ± 0.177	71.7
SCAUf127	22	a	1	1	A	*R. anhuiense* CCBAU 23252^T (98.9%)	0.733 ± 0.065	41.3
SCAUf128	9	a	1	1	A		0.707 ± 0.188	13.3
SCAUf100	10	e	13	21	M	*R. vallis* CCBAU 65647^T (99.0%)	0.689 ± 0.034	11.0
SCAUf104	24	f	16	23	I	*R. anhuiense* CCBAU 23252^T (99.2%)	0.680 ± 0.082	15.0
SCAUf89	20	a	10	2	B		0.680 ± 0.060	21.3
SCAUf119	4	a	1	1	A		0.667 ± 0.188	20.0
SCAUf135	4	a	1	1	A		0.663 ± 0.114	16.7
SCAUf105	3	a	3	11	A	*R. anhuiense* CCBAU 23252^T (99.3%)	0.660 ± 0.020	5.0
SCAUf108	3	a	7	15	A		0.650 ± 0.031	6.3
SCAUf116	25	a	1	1	A		0.647 ± 0.189	47.0
SCAUf112	21	a	1	1	A		0.647 ± 0.099	11.7
SCAUf96	5	a	1	1	A		0.643 ± 0.070	21.3
SCAUf88	19	a	2	8	A		0.632 ± 0.073	15.0
SCAUf144	4	b	24	18	C	*R. fabae* CCBAU 33202^T (99.9 %)	0.605 ± 0.090	24.0
SCAUf97	19	a	1	1	A		0.573 ± 0.032	31.0
SCAUf91	6	a	2	8	A	*R. anhuiense* CCBAU 23252^T (99.8%)	0.533 ± 0.087	13.0
SCAUf130	23	a	18	6	A		0.532± 0.086	23.0
SCAUf93	19	d	14	20	N	*A. radiobacter* NCPPB 2437^T (97.3%)	0.532± 0.069	0
SCAUf98	19	a	1	1	A		0.517± 0.059	0

[a] *CK: uninoculated treatment in the symbiotic efficiency test. Representative isolates for sequencing in bold.*
[b] *Sampling sites are the same as* ***Figure 1****.*
[c] *Genotype: the combination of the restriction patterns obtained by enzymes MspI, HaeIII, TaqI, and HinfI.*
[d] *CACAI: clustered analysis of combined ARDRA and IGS-RFLP, Groups were defined at 94.5% similarity level.*
[e] *MLSA, multilocus sequence analysis of combined recA, atpD, and glnII. The percentages are sequence similarities to the closely related species or the closest type strain. ^{T}Type strain.*
[f] *↑*Significantly higher shoot dry mass than that in CK treatment according to the LSD test (P = 0.05). Data presented as mean value ± standard deviation (n = 3, except n = 2 for SCAUf93 and SCAUf98).*

length polymorphism analysis. Primer pairs P1, P6 and pHr(F), p23SR01(R) (**Table 1**) were used for polymerase chain reaction (PCR) amplification. Amplification products (5 μl) were digested separately by four restriction enzymes *Hin*fI, *Taq*I, *Msp*I, and *Hae*III following the manufacturer's instructions (Fermentas, EU). The fragments were separated by gel electrophoreses in 2% agarose with 0.5 μg ml^{-1} ethidium bromide at 80 V for 3 h and photographed. Amplified ribosomal DNA restriction analysis (ARDRA) and IGS-RFLP were done by combining the results from the four restrictions. Clustered analysis of combined ARDRA and IGS-RFLP (CACAI) was conducted by UPGM clustering algorithm in the NTSYS program (Rohlf, 1990).

Sequencing of Housekeeping and Symbiotic Genes

According to the results of CACAI, representative strains were selected for sequencing of housekeeping and symbiotic genes. To facilitate the comparison of faba bean nodulating diversity in Panxi and other parts of Sichuan, we applied the same methods as in our earlier study on rhizobia from Sichuan hilly areas (Xu et al., 2015). 16S rDNA was amplified as described

TABLE 3 | The *nodC* and *nifH* types in the representative strains isolated from faba bean.

Representative strain	*nodC*[a]	*nifH*[a]
R. anhuiense SCAUf104	*R. fabae*	HRsp1
R. anhuiense SCAUf127	*R. fabae*	HRsp1
R. fabae SCAUf144	*R. fabae*	HRsp1
R. sophorae SCAUf86	*R. fabae*	HRsp2
R. anhuiense SCAUf131	*R. fabae*	HRsp2
R. anhuiense SCAUf140	*R. laguerreae*	HRsp1
R. anhuiense SCAUf105	CRla1	HRsp2
R. vallis SCAUf100	CRsp1	HRsp3
R. anhuiense SCAUf91	CRsp2	HRsp4
Rhizobium sp.I f99	CRsp2	HRsp4
Rhizobium sp.I SCAUf90	CRsp2	HRsp4
Rhizobium sp.I SCAUf133	CRsp2	HRsp4
Rhizobium sp.II SCAUf94	CRsp2	HRsp4
Rhizobium sp.II SCAUf106	CRsp2	HRsp4
Rhizobium sp.II SCAUf109	CRsp2	HRsp4

[a]*Closest type strain or clade.*

above. Three housekeeping genes *atpD*, *glnII,* and *recA* and two symbiotic genes *nifH* and *nodC* were amplified as described in **Table 1**. The PCR products were sequenced directly at BGI Tech (Shenzhen, China). Sequences have been deposited to NCBI (National Center for Biotechnology Information research database) nucleotide database under the accession numbers of KU947312-KU947400.

The sequences of the housekeeping and symbiotic genes were compared with sequences in NCBI, and the 16S rDNA sequences were compared with sequences in EzTaxon (http://www.ezbiocloud.net/) using BLASTN. Phylogenetic analyses of sequences from our isolates and reference sequences from databases were done using a Neighbor-Joining method in MEGA 6.0 (Tamura et al., 2013) with 1,000 bootstrapped replicates. Genospecies were defined by multilocus sequence analysis (MLSA) using concatenated sequence of three housekeeping genes applying 97% average nucleotide identity as the threshold (Cao et al., 2014).

RESULTS

Nodulation, Plant Growth Promoting Ability, and Genetic Diversity of Faba Bean Isolates

We isolated 65 strains from root nodules of faba bean growing in Panxi, China (**Table 2**). All but two of the strains formed nodules on the roots of faba bean with the average nodule numbers ranging from 3.0 to 98.5 per plant. No nodules were detected on the roots of the uninoculated plants (**Table 2**). The plant growth promoting ability of the isolates was assessed by measuring the dry masses of the inoculated plants. The eleven strains that significantly increased the plant shoot dry mass ($p < 0.05$) were considered as potential inoculant strains (**Table 3**).

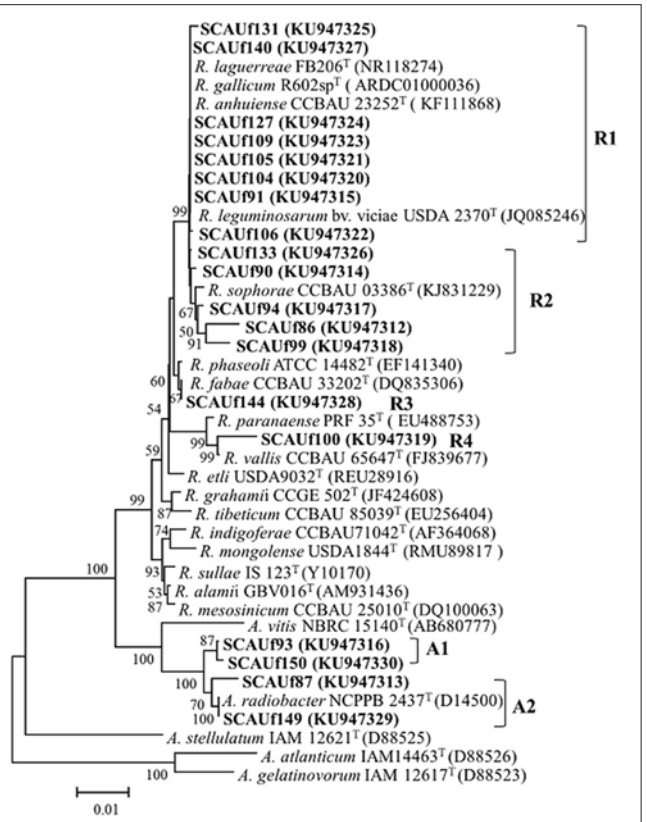


FIGURE 2 | Neighbor-joining tree based on 16S rDNA (1,336 nt) presenting the phylogenetic relationship among the representative strains isolated from faba bean (in bold) and reference strains. Bootstrap values ≥50% are shown on the branches. Genbank accession numbers are in parentheses. Scale bar = 1% substitutions per site. *R, Rhizobium*; *A, Agrobacterium*.

Amplification of the 16S rDNA gene resulted in an approximately 1,500 bp band from all the isolates. In the 16S rDNA PCR-RFLP, nine fragment pattern types (a-i) were observed: type a included 54 strains, types b, f, and h included two strains each, and types c, d, e, g, and i included one strain each (**Table 2**).

For the majority of strains, IGS PCR resulted in a single band ranging from 1,900 to 2,200 bp, whereas for strains SCAUf90 and SCAUf99 IGS PCR resulted in two and three bands, respectively (**Table 2**). The strains were divided to 25 IGS-RFLP types. In the combined analysis of 16S rDNA RFLP and IGS-RFLP (CACAI) the strains were divided into 14 CACAI groups at 94.5% similarity level and 26 CACAI genotypes (**Table 2**). CACAI group A was the largest group including 40 isolates with CACAI genotypes 1, 5, 6, 8, and 15. Seven of the plant growth promoting strains represented genotype 1, and the other four were assigned to genotypes 5, 9, 24, and 25.

16S rDNA Phylogeny

Based on CACAI groups as well as considering the sites of isolation of the strains, 19 representative strains were selected for sequencing. In the 16S rDNA phylogenetic tree

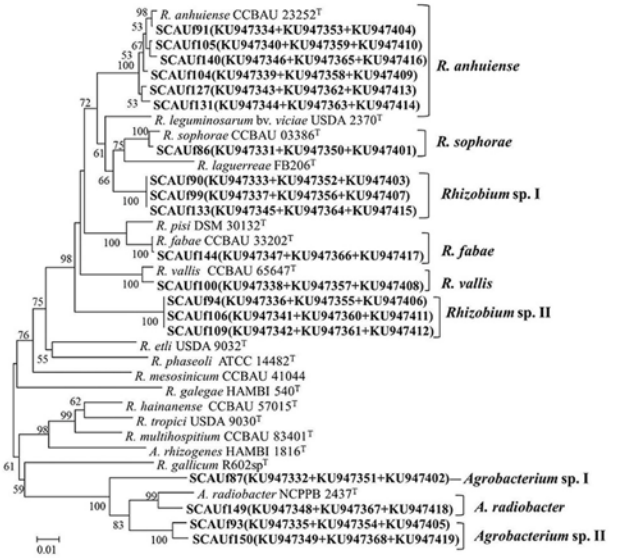


FIGURE 3 | Neighbor-joining tree based on multilocus sequence analysis using concatenated sequence of *atpD* (395 nt), *glnII* (483 nt), and *recA* (344 nt) genes presenting the phylogenetic relationship among the representative strains isolated from faba bean (in bold) and reference strains. Bootstrap values ≥50% are shown on the branches. Genbank accession numbers are in parentheses. Scale bar = 1% substitutions per site. *R, Rhizobium*, *A, Agrobacterium*.

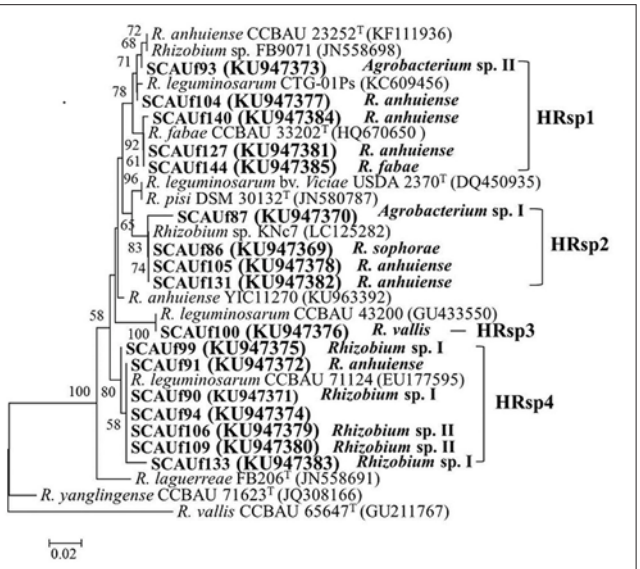


FIGURE 4 | Neighbor-joining tree based on *nifH* (329 nt) gene of 17 representative strains isolated from faba bean (in bold) and reference strains. Genbank accession numbers are in parentheses. Bootstrap values ≥50% are shown on the branches. Scale bar = 2% substitutions per site. *R, Rhizobium*.

(**Figure 2**), the strains clustered into six distinct clades with the reference strains. Four clades were related to *Rhizobium* (R group) and two to *Agrobacterium* (A group). Four strains clustered with *Agrobacterium radiobacter* type strain with 98.3–99.8% similarities. SCAUf144 clustered with *R. fabae* with 100% similarity, SCAUf100 clustered with *R. vallis* with 98.6% similarity, and SCAUf86, SCAUf90, SCAUf94, SCAUf99, and SCAUf133 clustered with *Rhizobium sophorae* into clade R2 with 97.9–99.9% similarities. The other eight strains clustered into a distinct clade with *R. gallicum*, *Rhizobium anhuiense*, *R. laguerreae,* and *Rhizobium leguminosarum* with similarities ranging from 99.8 to 100%.

Multilocus Sequence Analysis

In the multilocus sequence analysis (MLSA) based on housekeeping genes *atpD*, *glnII* and *recA*, the 19 representative strains clustered into nine distinct clades related to *Rhizobium* and *Agrobacterium* species (**Figure 3**). SCAUf86 was 99.3% similar to *R. sophorae* CCBAU 03386^T, thus assigned as *R. sophorae*. SCAUf90, SCAUf99 and SCAUf133 clustered separately and were assigned as *Rhizobium* sp. I, as did SCAUf94, SCAUf106 and SCAUf109 that were assigned as *Rhizobium* sp. II. SCAUf91, SCAUf104, SCAUf105, SCAUf127, SCAUf131, and SCAUf140 were 98.5–99.8% similar to *R. anhuiense* type strain, thus assigned as *R. anhuiense* strains. SCAUf144 and SCAUf100 clustered with *R. fabae* CCBAU 33202^T and *R. vallis* CCBAU 65647^T, respectively, thus assigned as *R. fabae* and *R. vallis*, respectively. As in 16S rDNA analysis, four strains clustered with *Agrobacterium* in the MLSA. Because no *glnII* sequences of the relevant *Agrobacterium* type strains except *A. radiobacter* type strain were available in the GenBank sequence database, the relationships between *Agrobacterium* strains were studied based on non-type strains (Supplementary Figure S1). SCAUf87 clustered separately, and was assigned as *Agrobacterium* sp. I. SCAUf93 and SCAUf150 clustered separately and were assigned as *Agrobacterium* sp. II. SCAUf149 was 97.3% similar to *A. radiobacter* NCPPB 2437^T with, thus assigned as *A. radiobacter*.

Diversity of Symbiosis Genes

For both *nifH* and *nodC* amplification was not successful with one primer pair only, possibly due to differences in primer binding sites. Approximately 700 bp fragments were obtained using primer pair nifHctg/nifHI (13 representative strains), and 400 bp products using primer pair nifH1F/ nifH1R (SCAUf87, SCAUf105, SCAUf133, SCAUf140). Amplification of *nifH* from SCAUf149 and SCAUf150 was not successful. Seventeen strains clustered into four clades in the *nifH* phylogenetic tree (**Figure 4**, **Table 3**). The *nifH* of *Agrobacterium* sp. II SCAUf93 and *R. anhuiense* SCAUf104 were 99.7 and 99.6%, respectively, similar to that of *R. anhuiense* CCBAU 23252^T. The *nifH* of *R. anhuiense* SCAUf127 and *R. fabae* SCAUf144 were 100% similar to that of *R. fabae* type strain. The *nifH* of *R. sophorae* SCAUf86, *R. anhuiense* SCAUf105 and SCAUf131 clustered with that of *R. leguminosarum* USDA 2370^T with 98.4% similarity. *R. vallis* SCAUf100 clustered with *R. leguminosarum* CCBAU 43200 with 100% similarity. The strains *R. anhuiense* SCAUf91, *Rhizobium* sp. I SCAUf90, *Rhizobium* sp. II SCAUf94, SCAUf106 and SCAUf109 carried *nifH* 100% similar to that of *R. leguminosarum* CCBAU 71124.

Nearly 600 bp *nodC* fragments were amplified from thirteen representative strains using primer pair nodC540/nodC1160. Amplification from strains SCAUf105 and SCAUf140 was successful only while using *R. leguminosarum* sv. *viciae nodC* specific primer pair nodCf/nodCr. Amplification of *nodC* from strains assigned as *Agrobacterium* was not successful. Fifteen strains clustered into five clades in the *nodC* phylogenetic tree (**Figure 5**, **Table 3**). The *nodC* of *R. anhuiense* SCAUf131, SCAUf127, and SCAUf104, *R. fabae* SCAUf144 and *R. sophorae* SCAUf86 were 100% similar to that of *R. fabae* type strain (**Figure 5A**). Similarly to the *nifH* analysis, the *nodC* of *R. vallis* SCAUf100 clustered with *nodC* from non-type strains. Seven strains carried *nodC* 100% similar to that of *R. leguminosarum* non-type strain. The *nodC* of *R. anhuiense* SCAUf140 and SCAUf105 (**Figure 5B**) were 100 and 97.2% similar to that of *R. laguerreae*.

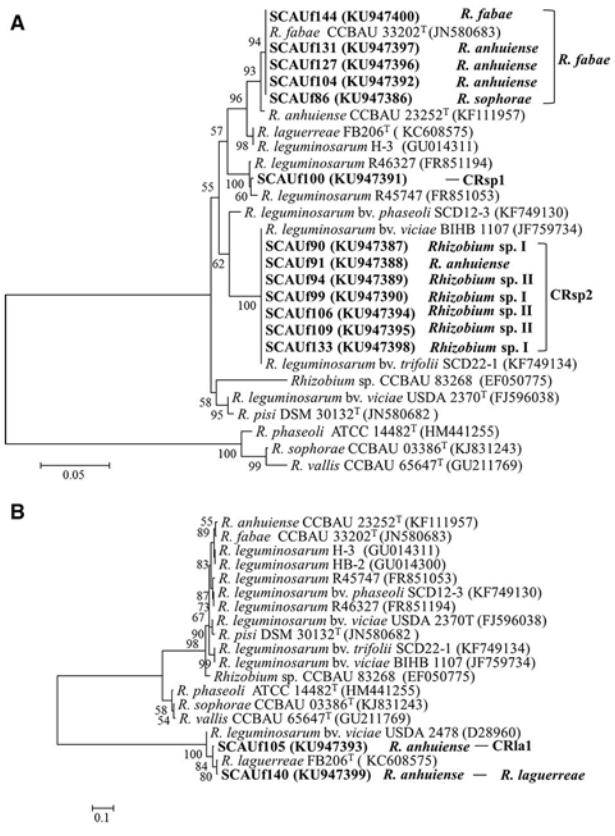


FIGURE 5 | Neighbor-joining tree based on *nodC* gene of 13 representative strains (499 nt) **(A)** and 2 representative strains (230 nt) **(B)** isolated from faba bean (in bold) and reference strains. Genbank accession numbers are in parentheses. Bootstrap values ≥50% are shown on the branches. Scale bar = 5% substitutions per site. *R, Rhizobium*.

DISCUSSION

Due to overcutting and mining Panxi in Southwestern China has suffered serious soil degradation and heavy metal contamination (Xu et al., 2013; Yu et al., 2014). Reclaiming the soils requires sustainable and efficient yet low economic input methods, for example utilization of biological nitrogen fixation (BNF) by legume-rhizobium symbiosis instead of relatively cheap nitrogen fertilizer. To facilitate the utilization of BNF we tested the plant growth promoting ability of rhizobial isolates from faba bean in search of locally adapted, potential inoculant strains. In Ethiopia, chickpea nodulating rhizobia showed big differences in efficiency and nodule numbers, and strains with similar efficiencies did not necessarily induce similar numbers of nodules and vice versa (Tena et al., 2017). Similarly, in our study variations in plant growth promoting ability and nodule numbers were large, and only 11 strains increased faba bean shoot dry mass significantly. Similar to *Leucaena leucocephala* isolates from Panxi (Xu et al., 2013), for many of the strains inoculation resulted in dry mass lower than that in uninoculated plants, highlighting the need to apply selected inocula to promote BNF.

The diversity and identity of the strains were assessed using molecular methods. Faba bean is nodulated by *R. fabae*, *R. leguminosarum*, *R. anhuiense, R. laguerraeae* and *A. radiobacter* strains, and the dominant species is different in different regions (Tian et al., 2007, 2008; Youseif et al., 2014; Xu et al., 2015; Zhang et al., 2015; Xiong et al., 2017). In our study, the isolated

TABLE 4 | Rhizobial species, nodD and nodC diversity in subtropical China.

Province	Climate	Species	*nodD* types	*nodC* types	Dominant species	References
Yunnan	Subtropical highland, humid subtropical	5[a]	3[c]	nd	*R. laguerraeae*	Xiong et al., 2017
Yunnan	Subtropical highland, humid subtropical	1[b]	4[d]	1[d]	*R. leguminosarum*	Tian et al., 2007
Anhui, Jiangxi, Henan, Zhejiang	Humid subtropical	3[a]	5[c]	nd	*R. anhuiense*	Xiong et al., 2017
Anhui, Jiangxi	Humid subtropical	2[b]	3[d]	1[d]	*R. leguminosarum*	Tian et al., 2007
Sichuan	Humid subtropical	5[b]	nd	3[e]	*R. leguminosarum*	Xu et al., 2015
Panxi	See Introduction	6[b]	nd	5[e]	*R. anhuiense*	This study

Nd, not determined.
[a]*Determined by amplicon sequencing targeting rpoB.*
[b]*Determined by multilocus sequence analysis.*
[c]*Determined by amplicon sequencing targeting nodD.*
[d]*Determined by RFLP.*
[e]*Determined by Sanger sequencing nodC.*

strains were related to genera *Rhizobium* and *Agrobacterium*. The *Rhizobium* strains were assigned as representing *R. anhuiense*, *R. fabae, R. sophorae*, and *R. vallis*, and two putative new species in the genus *Rhizobium*. Similar to subtropical provinces in East China (Xiong et al., 2017), *R. anhuiense* was the dominant species among the faba bean nodulating rhizobia in Panxi. To our knowledge, *R. sophorae*, a symbiont of medicinal legume *Sophora flavescens* (Jiao et al., 2015) and *R. vallis*, a symbiont of *Phaseolus vulgaris* (Wang et al., 2011), have not earlier been shown to nodulate faba bean.

The rhizobia-legume symbiosis benefits sustainable agriculture due to the symbiotic nitrogen fixation capacity that needs two key points: nodule infection and nitrogen fixation, both of which need the regulation of symbiosis related genes (Masson-Boivin et al., 2009). In the present study, *nifH* gene that is the structural gene encoding the nitrogenase Fe protein (Masson-Boivin et al., 2009), and *nodC* that is the gene encoding enzymes involved in the synthesis of the core structure of the Nod-factor (Geremia et al., 1994) were selected for sequencing to analyze the symbiotic phylogeny of the faba bean rhizobia in Panxi region. The symbiotic genes are commonly located on a symbiotic plasmid or island which may be transferred (Laranjo et al., 2012; Bakhoum et al., 2014). Faba bean nodulating *R. leguminosarum* strains that carried four different types of nodulation gene *nodD* had all similar *nodC* (**Table 4**) (Tian et al., 2007). The five types of *nodC* detected in this study suggest higher diversity at symbiosis related gene level. However, considering the six different *nifH-nodC* combinations in our study, the faba bean isolates from Yunnan (Tian et al., 2007) and Panxi were approximately equally diverse.

The *Desmodium* nodulating rhizobium strains in Panxi region were quite different from those in other places such as temperate and subtropical region of China and Central and North America, possibly due to the special environmental conditions (Xu et al., 2016). The faba bean rhizobia in this area were approximately as diverse as in Sichuan hilly areas and in Yunnan (**Table 4**) (Xu et al., 2015; Xiong et al., 2017), yet more diverse than in other parts of subtropical China (Tian et al., 2007; Xiong et al., 2017). When looking at the species-*nifH-nodC* combinations it is noteworthy that all but two of the six *R. anhuiense* isolates were different. The symbiosis and nitrogen fixation related genes of rhizobia can be transferred laterally (Sullivan et al., 1995). However, whether the increase in diversity in Panxi was caused by lateral transfer cannot be concluded based on our data.

In conclusion, eleven out of 65 faba bean strains in Panxi area could significantly promote plant growth, and were thus considered as potential inoculants. The nodule isolates in this area were diverse belonging to nine species. *R. anhuiense*, the dominant faba bean nodulating species in this area, was diverse both at plant growth promoting ability and symbiosis related gene levels.

AUTHOR CONTRIBUTIONS

KX and YC conceived and designed the experiments. KX supervised the experiments. YC, LZ, PP, and KX contributed to discussion of the results, and writing and revising the manuscript. LZ performed most of the experiments and analyzed data. QC and CW participated in collecting faba bean nodules and relevant soil information, and relevant meteorological information from the Sichuan meteorological bureau. QL created the map in **Figure 1** and revised the manuscript. All authors contributed to writing the article.

FUNDING

This work was supported by the National Key Research and Development Program of China (2016YFD0300300).

ACKNOWLEDGMENTS

We would like to thank Mr. Zhi Heng Liu for collecting faba bean nodules in Panxi.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.01338/full#supplementary-material

REFERENCES

Bakhoum, N., Galiana, A., Le Roux, C., Kane, A., Duponnois, R., Ndoye, F., et al. (2014). Phylogeny of nodulation genes and symbiotic diversity of *Acacia senegal* (L.) Willd. and A. seyal (Del.) Mesorhizobium strains from different regions of Senegal. *Microb. Ecol.* 69, 641–651. doi: 10.100/7/s00248-014-0507-1

Cao, Y., Wang, E. T., Zhao, L., Chen, W. M., and Wei, G. H. (2014). Diversity and distribution of rhizobia nodulated with *Phaseolus vulgaris* in two ecoregions of China. *Soil Biol. Biochem.* 78, 128–137. doi: 10.1016/j.soilbio.2014.07.026

Chen, Q., Chen, W. X., Zhang, X. P., Li, D. Y., and Lindstrom, K. (2004). Genetic diversity of rhizobia isolated from *Pueraria* spp. in Sichuan, China. *Sci. Agric. Sin.* 37, 1641–1646. doi: 10.3321/j.issn:0578-1752.2004.11.010

Crews, T. E., and Peoples, M. B. (2004). Legume versus fertilizer sources of nitrogen: ecological tradeoffs and human needs. *Agr. Ecosyst. Environ.* 102, 279–297. doi: 10.1016/j.agee.2003.09.018

Gentzbittel, L., Andersen, S. U., Ben, C., Rickauer, M., Stougaard, J., and Young, N. D. (2015). Naturally occurring diversity helps to reveal genes of adaptive importance in legumes. *Front. Plant Sci.* 6:269. doi: 10.3389/fpls.2015.00269

Geremia, R. A., Mergaert, P., Geelen, D., Van Montagu, M., and Holsters, M. (1994). The NodC protein of *Azorhizobium caulinodans* is an N-acetylglucosaminyltransferase. *Proc. Natl. acad. Sci. U.S.A.* 91, 2669–2673.

Graham, P. H., and Vance, C. P. (2003). Legumes: importance and constraints to greater use. *Plant physiol.* 131, 872–877. doi: 10.1104/pp.017004

Gurkanli, C. T., Ozkoc, I., and Gunduz, I. (2014). Genetic diversity of *Vicia faba* L. and *Pisum sativum* L. nodulating rhizobia in the central Black Sea region of Turkey. *Ann. Microbiol.* 64, 99–112. doi: 10.1007/s13213-013-0638-5

Haciseferoğullar,i, H., Gezer, I., Bahtiyarca, Y., and Menge,ş, H. O. (2003). Determination of some chemical and physical properties of Sakiz faba bean (*Vicia faba* L. Var. major). *J. Food Eng.* 60, 475–479. doi: 10.1016/S026 0-8774(03)00075-X

Jiao, Y. S., Yan, H., Ji, Z. J., Liu, Y. H., Sui, X. H., Wang, E. T., et al. (2015). *Rhizobium sophorae* sp. nov. and *Rhizobium sophoriradicis* sp. nov., nitrogen-fixing rhizobial symbionts of the medicinal legume *Sophora flavescens*. *Int. J. Syst. Evol. Microbiol.* 65(Pt 2), 497–503. doi: 10.1099/ijs.0.068916-0068916-0

Köpke, U., and Nemecek, T. (2010). Ecological services of faba bean. *Field Crops Res.* 115, 217–233. doi: 10.1016/j.fcr.2009.10.012

Laguerre, G., Nour, S. M., Macheret, V., Sanjuan, J., Drouin, P., and Amarger, N. (2001). Classification of rhizobia based on *nodC* and *nifH* gene analysis reveals a close phylogenetic relationship among *Phaseolus vulgaris* symbionts. *Microbiology* 147, 981–993. doi: 10.1099/00221287-147-4-981

Laranjo, M., Young, J. P. W., and Oliveira, S. (2012). Multilocus sequence analysis reveals multiple symbiovars within *Mesorhizobium* species. *Syst. Appl. Microbiol.* 35, 359–367. doi: 10.1016/j.syapm.2 012.06.002

Li, L., Yang, T., Redden, R., He, W., and Zong, X. (2016). Soil fertility map for food legumes production areas in China. *Sci. Rep.* 6:26102. doi:10.1038/srep26102

Massol-Deya, A. A., Odelson, D. A., Hickey, R. F., and Tiedje, J. M. (1995). "Bacterial community fingerprinting of amplified 16S and 16–23S ribosomal DNA gene sequences and restriction endonuclease analysis (ARDRA)," in *the Molecular Microbial Ecology Manual*, eds A. D. L. Akkermans, J. D. Van Elsas, and F. J. De Bruijn (Dordrecht: Springer), 289–296.

Masson-Boivin, C., Girau, D. E., Perret, X., and Batut, J. (2009). Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? *Trends Microbiol.* 17, 458–466. doi: 10.1016/j.tim.2009.07.004

Mei, P. P., Gui, L. G., Wang, P., Huang, J. C., Long, H. Y., Christie, P., et al. (2012). Maize/faba bean intercropping with rhizobia inoculation enhances productivity and recovery of fertilizer P in a reclaimed desert soil. *Field Crops Res.* 130, 19–27. doi: 10.1016/j.fcr.2012.02.007

Moschetti, G., Peluso, A., Protopapa, A., Anastasio, M., Pepe, O., and Defez, R. (2005). Use of nodulation pattern, stress tolerance, *nodC* gene amplification, RAPD-PCR and RFLP−16S rDNA analysis to discriminate genotypes of *Rhizobium leguminosarum* biovar viciae. *Syst. Appl. Microbiol.* 28, 619–631. doi: 10.1 016/j.syapm.2005.03.009

Navarro, E., Simonet, P., Normand, P., and Bardin, R. (1992). Characterization of natural populations of *Nitrobacter* spp. using PCR/RFLP analysis of the ribosomal intergenic spacer. *Arch. Microbiol.* 157, 107–115

Rivas, R., Velázquez, E., Willems, A., Vizcaíno, N., Subba-Rao, N. S., Mateos, P. F., et al. (2002). A new species of *Devosia* that forms a unique nitrogen-fixing root-nodule symbiosis with the aquatic legume *Neptunia natans* (Lf) Druce. *Appl. Environ. Microbiol.* 68, 5217–5222. doi: 10.1128/AEM.68.11.5217-5222.2002

Rohlf, F. J. (1990). *NTSYS-pc, Numerical taxonomy and Multivariate Analysis System, Version 2.02*. New York, NY: Exeter Software.

Sarita, S., Sharma, P. K., Priefer, U. B., and Prell, J. (2005). Direct amplification of rhizobial *nodC* sequences from soil total DNA and comparison to *nodC* diversity of root nodule isolates. *FEMS Microbiol. Ecol.* 54, 1–11. doi: 10.1016/j.femsec.2005.02.015

Song, Y. N., Zhang, F. S., Marschner, P., Fan, F. L., Gao, H. M., Bao, X. G., et al. (2007). Effect of intercropping on crop yield and chemical and microbiological properties in rhizosphere of wheat (*Triticum aestivum* L.), maize (*Zea mays* L.), and faba bean (*Vicia faba* L.). *Biol. Fertil. Soils* 43, 565–574. doi: 10.1007/s00374-006-0139-9

Sullivan, J. T., Patrick, H. N., Lowther, W. L., Scott, D. B., and Ronson, C. W. (1995). Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8985–8989. doi: 10.1073/pnas.92.19.8985

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tan, Z. Y., Xu, X. D., Wang, E. T., Gao, J. L., Martinez-Romero, E., and Chen, W. X. (1997). Phylogenetic and genetic relationships of *Mesorhizobium tianshanense* and related rhizobia. *Int J. Syst. Bacteriol.* 47, 874–879. doi: 10.1099/00207713-47-3-874

Tena, W., Wolde-Meskel, E., Degefu, T., and Walley, F. (2017). Genetic and phenotypic diversity of rhizobia nodulating chickpea (*Cicer arietinum* L.) in soils from southern and central Ethiopia. *Can. J. Microbiol.* 63, 690–707. doi: 10.1139/cjm-2016-0776

Terefework, Z., Kaijalainen, S., and Lindström, K. (2001). AFLP fingerprinting as a tool to study the genetic diversity of *Rhizobium galegae* isolated from *Galega orientalis* and *Galega officinalis*. *J. Biotechnol.* 91, 169–180. doi: 10.1016/S0168-1656(01)00338-8

Thilakarathna, M. S., and Raizada, M. N. (2017). A meta-analysis of the effectiveness of diverse rhizobia inoculants on soybean traits under field conditions. *Soil. Biol. Biochem.* 105, 177–196. doi: 10.1016/j.soilbio.2016.11.022

Tian, C. F., Wang, E. T., Han, T. X., Sui, X. H., and Chen, W. X. (2007). Genetic diversity of rhizobia associated with *Vicia faba* in three ecological regions of China. *Arch. Microbiol.* 188, 273–282. doi: 10.1007/s00203-007-0245-6

Tian, C. F., Wang, E. T., Wu, L. J., Han, T. X., Chen, W. F., Gu, C. T., et al. (2008). *Rhizobium fabae* sp. nov., a bacterium that nodulates *Vicia faba*. *Int. J. Syst. Evol. Microbiol.* 58, 2871–2875. doi: 10.1099/ijs.0.2008/000703-0

Vincent, J. M. (1970). *A Manual for the Practical Study of Root Nodule Bacteria. International Biological Programme Handbook no. 15*. Oxford: Blackwell Scientific Publications.

Vinuesa, P., Silva, C., Werner, D., and Martínez-Romero, E. (2005). Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in *Bradyrhizobium* species cohesion and delineation, *Mol. Phylogenet. Evol.* 34, 29–54. doi: 10.1016/j.ympev.2004.08.020

Wang, F. Q., Wang, E. T., Zhang, Y. F., and Chen, W. X. (2006). Characterization of rhizobia isolated from *Albizia* spp. in comparison with microsymbionts of *Acacia* spp. and *Leucaena leucocephala* grown in China. *Syst. Appl. Microbiol.* 29, 502–517. doi:10.1016/j.syapm.2005.12.010

Wang, F., Wang, E. T., Wu, L. J., Sui, X. H., Li, Y., and Chen, W. X. (2011). *Rhizobium vallis* sp. nov., isolated from nodules of three leguminous species. *Int. J. Syst. Evol. Microbiol.* 61, 2582–2588. doi: 10.1099/ijs.0.026484-0

Wu, B., Zhu, C., Li, D., Dong, K., Wang, X., and Shi, P. (2006). Setting biodiversity conservation priorities in the Forests of the Upper Yangtze Ecoregion based on ecoregion conservation methodology. *Biodiver. Sci.* 14, 87–97. doi: 10.1360/biodiv.050232

Xie, H., Zhang, X. P., and Lindstrom, K. (2009). RFLP genetic diversity of rhizobia isolated from *Pueraria lobata* in Panxi. *J. Sichuan Agric. Univ.* 27, 314–320. doi: 10.3969/j.issn.1000-2650.2009.03.010

Xiong, H. Y., Zhang, X. X., Guo, H. J., Ji, Y. Y., Li, Y., Wang, X. L., et al. (2017). The epidemicity of facultative microsymbionts in faba bean rhizosphere soils. *Soil. Biol. Biochem.* 115, 243–252. doi: 10.1016/j.soilbio.2017.08.032

Xu, K. W., Penttinen, P., Chen, Y. X., Chen, Q., and Zhang, X. (2013). Symbiotic efficiency and phylogeny of the rhizobia isolated from *Leucaena leucocephala* in arid-hot river valley area in Panxi, Sichuan, China. *Appl. Microbiol. Biotechnol.* 97, 783–793. doi: 10.1007/s00253-012-4246-2

Xu, K. W., Zou, L., Penttinen, P., Wang, K., Heng, N. N., Zhang, X. P., et al. (2015). Symbiotic effectiveness and phylogeny of rhizobia isolated from faba bean (*Vicia faba* L.) in Sichuan hilly areas, China. *Syst.Appl. Microbiol.* 38, 515–523. doi: 10.1016/j.sya pm.2015.06.009

Xu, K. W., Zou, L., Penttinen, P., Zeng, X., Liu, M., and Zhao, K., et al (2016). Diversity and phylogeny of rhizobia associated with *Desmodium* spp. in Panxi, Sichuan, China. *Syst. Appl. Microbiol.* 39, 33–40. doi: 10.1016/j.syapm.2015.10.005

Youseif, S. H., El-Megeed, F. H. A., Ageez, A., Cocking, E. C., and Saleh, S. A. (2014). Phylogenetic multilocus sequence analysis of native rhizobia nodulating faba bean (*Vicia faba* L.) in Egypt. *Syst. Appl. Microbiol.* 37, 560–569. doi: 10.1016/j.syapm.2014.10.001

Yu, X., Li, Y., Zhang, C., Liu, H., Liu, J., and Zheng, W., et al (2014). Culturable heavy metal-resistant and plant growth promoting bacteria in V-Ti magnetite mine tailing soil from Panzhihua, China. *PloS ONE* 9:e106618. doi: 10.1371/journal.pone.0106618

Zhang, Y. J., Zheng, W. T., Everall, I., Young, J. P. W., Zhang, X. X., Tian, C. F., et al. (2015). *Rhizobium anhuiense* sp. nov., isolated from effective nodules of *Vicia faba* and *Pisum sativum*. *Int. J. Syst. Evol. Microbiol.* 65, 2960–2967. doi: 10.1099/ijs.0.000365

Conflict of Interest Statement: 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.

Copyright © 2018 Chen, Zou, Penttinen, Chen, Li, Wang and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 22 June 2018
doi: 10.3389/fmicb.2018.01336

Metformin Alters Gut Microbiota of Healthy Mice: Implication for Its Potential Role in Gut Microbiota Homeostasis

Wei Ma[1,2,3†], Ji Chen[2,4†], Yuhong Meng[2,4†], Jichun Yang[2,4], Qinghua Cui[1,2,4] and Yuan Zhou[1,2*]*

[1] Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University, Beijing, China, [2] Ministry of Education Key Laboratory of Molecular Cardiovascular Sciences, Peking University, Beijing, China, [3] Central Laboratory, PLA Navy General Hospital, Beijing, China, [4] Department of Physiology and Pathophysiology, School of Basic Medical Sciences, Peking University, Beijing, China

In recent years, the first-line anti-diabetic drug metformin has been shown to be also useful for the treatment of other diseases like cancer. To date, few reports were about the impact of metformin on gut microbiota. To fully understand the mechanism of action of metformin in treating diseases other than diabetes, it is especially important to investigate the impact of long-term metformin treatment on the gut microbiome in non-diabetic status. In this study, we treated healthy mice with metformin for 30 days, and observed 46 significantly changed gut microbes by using the 16S rRNA-based microbiome profiling technique. We found that microbes from the *Verrucomicrobiaceae* and *Prevotellaceae* classes were enriched, while those from *Lachnospiraceae* and *Rhodobacteraceae* were depleted. We further compared the altered microbiome profile with the profiles under various disease conditions using our recently developed comparative microbiome tool known as MicroPattern. Interestingly, the treatment of diabetes patients with metformin positively correlates with colon cancer and type 1 diabetes, indicating a confounding effect on the gut microbiome in patients with diabetes. However, the treatment of healthy mice with metformin exhibits a negative correlation with multiple inflammatory diseases, indicating a protective anti-inflammatory role of metformin in non-diabetes status. This result underscores the potential effect of metformin on gut microbiome homeostasis, which may contribute to the treatment of non-diabetic diseases.

Keywords: metformin, gut microbiome, diabetes, 16S rRNA sequencing, MicroPattern

OPEN ACCESS

Edited by:
Xing Chen,
China University of Mining and Technology, China

Reviewed by:
Feng-Biao Guo,
University of Electronic Science and Technology of China, China
Chunyu Zhu,
Liaoning University, China

****Correspondence:***
Qinghua Cui
cuiqinghua@hsc.pku.edu.cn
Yuan Zhou
soontide6825@163.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: *07 April 2018*
Accepted: *31 May 2018*
Published: *22 June 2018*

Citation:
Ma W, Chen J, Meng Y, Yang J, Cui Q and Zhou Y (2018) Metformin Alters Gut Microbiota of Healthy Mice: Implication for Its Potential Role in Gut Microbiota Homeostasis. Front. Microbiol. 9:1336. doi: 10.3389/fmicb.2018.01336

INTRODUCTION

Metformin, also known as 1,1-dimethylbiguanide has been widely used in the treatment of type 2 diabetes mellitus (T2DM) since 1958 in United Kingdom and 1995 in United States (Witters, 2001). The main mechanisms underlying its anti-hyperglycemia effect include decreasing intestinal absorption of glucose, increasing insulin sensitivity and decreasing hepatic glucose production (Bailey and Turner, 1996; Hundal et al., 2000), which together result in reduction of basal and postprandial glucose levels. Because clinical investigations have shown that metformin has low

risk of hypoglycemia, modest weight loss, persistent anti-hyperglycemic effect and cardiovascular safety, it is now approved as one of the first-line drugs for treating T2DM (Garber et al., 2013).

Interestingly, it was shown in 2005 that metformin has the anti-cancer properties (Evans et al., 2005). Then a series of studies provided supporting evidence of its anti-cancer effects in a variety of cancer types such as ovarian cancer (Shank et al., 2012), endometrial cancer (Shafiee et al., 2014), breast cancer (Zhang and Li, 2014), liver cancer, pancreatic cancer, esophageal cancer, gastric cancer, and colorectal cancer (Franciosi et al., 2013). Metformin has been shown to reduce the incidence and mortality of cancer and block migration and invasion of tumor cells (Bao et al., 2012; Wu et al., 2015). Current knowledge pertaining to the molecular mechanisms underlying the anti-cancer activity of metformin is focused on two pathways that inhibit mTOR: (1) the AMPK-dependent pathway, in which metformin activates LKB1-AMPK to inhibit mTOR; (2) the AMPK-independent pathway, in which metformin inhibits mTOR via the PI3K/Akt/mTOR cascade (Sosnicki et al., 2016; Zhang and Guo, 2016). Besides, many studies have shown that metformin could also be utilized for the treatment of other diseases including obesity, polycystic ovary syndrome and tuberculosis (Gundelach et al., 2016; Igel et al., 2016; Restrepo, 2016). Moreover, it was reported that metformin also has potentially anti-aging effects (Novelle et al., 2016).

Body microbiota play extensive roles in physiology and it is therefore known as a "forgotten organ" (O'Hara and Shanahan, 2006). A lot of diseases are associated with gut microbiota, including cancer (Schwabe and Jobin, 2013; May et al., 2016), cardiovascular diseases (Koeth et al., 2013; Tang et al., 2013), obesity (Ley, 2010), diabetes (Wen et al., 2008; Qin et al., 2012), multiple sclerosis (Ezendam et al., 2008; Berer et al., 2011; Lee et al., 2011), neuromyelitis optica (Varrin-Doyer et al., 2012; Banati et al., 2013), Guillain–Barré syndrome (Ochoa-Reparaz et al., 2011), central nervous system disorders (Wang and Kasper, 2014) and autoinflammatory diseases (Lukens et al., 2014). For example, in obese individuals, *Bacteroidetes* is decreased whereas *Firmicutes* is increased (Furet et al., 2010), while *Prevotellaceae* that could produce H_2 is increased and methanogenic archaea which could utilize H_2 is also increased (Zhang et al., 2009). The co-existence of H_2-producing bacteria with relatively high numbers of H_2-utilizing methanogenic archaea in the gastrointestinal tract of obese individuals implies the plausible inter-species H_2 transfer between bacterial and archaeal species as an important mechanism for increasing intestinal energy uptake in obese persons. Moreover, in the gut of human subjects with type 2 diabetes, *Firmicutes* is significantly decreased (Larsen et al., 2010). The abundance of *Faecalibacterium prausnitzii* was negatively correlated with both diabetic and inflammatory markers, which indicated that *Faecalibacterium prausnitzii* could regulate inflammation in gut in diabetic patients (Furet et al., 2010). Finally, Bolte et al. found that in gut of autistic disorder, *Clostridium tetani* is increased while Finegold et al. found that *Bacteroidetes* is also increased. (Bolte, 1998; Finegold et al., 2010). Since the gut microbiota play important roles in the development of various diseases, they are intuitively one of the first targets for drugs and may also contribute to the effect of metformin in treating diseases like cancer and T2DM. Nevertheless, to date, few studies have taken gut microbiota into consideration. As a result, current knowledge about the mechanism of action of metformin is still not completed. Since metformin has been suggested to treat a wide spectrum of diseases other than T2DM even for healthy individuals, it is especially important to assess the effect of metformin on the gut microbiota considering its potential long-term usage in healthy conditions. What microbes are regulated by metformin in healthy individuals? How do these regulated microbes associate with other diseases? What is the difference of changes of gut microbiota between healthy and various diseases when treated with metformin? A profiling of the gut microbiome under healthy condition is necessary to answer these questions.

In this study, we treated healthy mice with metformin for 30 days and used 16S rRNA sequencing to evaluate the abundance of microbes in fecal samples. By comparing the metformin-treated healthy mice to the mock controls, we observed 46 significantly changed microbes. In addition, from previous publications, we also obtained significantly changed microbes from T2DM patients after metformin treatment (Forslund et al., 2015; Wu et al., 2017; Allin et al., 2018). We then used MicroPattern, a tool we recently developed for the comparison of microbiome profiles under different situations, to analyze these significantly altered microbes. By procedure, whether metformin could elicit different alterations of gut microbiota under diabetes and non-diabetic conditions, respectively, were evaluated and discussed.

MATERIALS AND METHODS

Animal Protocol and Sample Collection

This study was carried out in accordance with the principles of the Basel Declaration and recommendations of the Guide for the Care and Use of Laboratory Animals, US National Institutes of Health (NIH Publication No. 85-23, revised 1996). The protocol was approved by the Animal Research Committee of the Peking University Health Science Center. More specifically, 19 C57BL/6 healthy mice were separated into two groups: 9 were controls and 10 were included in the metformin-treated group. Until 8 weeks of age, mice were maintained on a chow diet. Then, mice in the metformin-treated group were treated with metformin (300 mg/kg of body weight) once daily via intragastric administration for 30 days. Mice in the control group were treated with an equivalent amount of saline via intragastric administration for 30 days. Fecal samples were obtained from 19 mice under sterile conditions. Because the microbiome profiling technique requires a large amount of raw materials, fecal was collected from 3 or 4 mice per sample (e.g., the fecal from the first, second and third mice was gathered together as the first sample). Finally, we acquired 3 metformin-treated samples and 3 mock control samples. Every sample was stored in a sterile 1.5 ml centrifuge tube at −80°C until microbiome profiling analysis.

16S rRNA Gene Sequencing

Microbial DNA was extracted from the fecal samples and the 16S rRNA gene of the isolated DNA was sequenced using Illumina Miseq2500 platform (service provided by GENE DENOVO Corporation) following the manufacturer's guidelines. 16S rRNA gene sequences de-multiplexing, quality control and operational taxonomic unit (OTU) binning were performed using Mothur version 1.3.4.0 with the standard pipeline (Schloss et al., 2009). The statistical test was performed in R, a free tool for scientific computing. OTU pathway analysis was performed using the Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) tool (Langille et al., 2013).

Enrichment Analysis and Disease Similarity Calculation

Enrichment analysis and disease similarity calculation were performed using MicroPattern (Ma et al., 2017). We just keep the significantly changed microbes at genus or species level as the input. At last, we acquired 46 microbes that changed significantly in metformin treated mice versus mock controls. To analyze the effect of metformin more comprehensively, we used the published human gut microbiome profiles under different diabetes situation, with or without metformin treatment (Forslund et al., 2015; Wu et al., 2017; Allin et al., 2018). First, Forslund et al. (2015) studied the effects of type 2 diabetes and metformin on the human gut microbiota. In their research, there are four group including healthy controls, type 1 diabetes mellitus (T1DM) patients, T2DM patients and metformin treated T2DM patients (MTT2DM). We acquired 36 significantly changed microbes from metformin treated T2DM patients contrasted against T2DM patients without metformin treatment, 26 significantly changed microbes from T2DM contrasted against healthy and 9 significantly changed microbes from T1DM patients contrasted against healthy controls. Second, Allin et al. (2018) studied the aberrant intestinal microbiota in individuals with prediabetes and 5 significantly changed microbes in prediabetes individuals versus healthy controls were obtained. Third, we got 29 significantly changed microbes from Wu et al.'s study about alteration of gut microbiome in treatment-naive T2DM patients after metformin treatment (Wu et al., 2017). We integrated those data and our data together for the comparison analysis. Finally, the significance of the similarity in microbiota changes was evaluated by permutation-based resampling test. More specifically, we shuffled the de-regulated microbiota between different diseases and re-calculated the similarity scores based on the randomly permutated microbiota. This procedure was repeated for 10000 times. For the observed positive similarity, if no higher similarity could be observed in more than 9000 out of 10000 such permutation tests, this similarity was considered significant. Likewise, for the observed negative similarity, if no lower similarity could be observed in more than 9000 out of 10000 such permutation tests, this similarity was considered significant. Such threshold also corresponded to a false discovery rate (FDR) threshold of 0.1.

RESULTS

Effect of Metformin of Mice Gut Microbiota

We treated healthy mice with metformin for 30 days and acquired significantly changed microbes in comparison with saline treated mock controls. To reduce noise, we just used the microbes whose tags occupy at least 0.1% of all tags. There is no significant difference of bacterial diversity between two groups with respect to Shannon diversity metrics (6.95 versus 6.87 for mean Shannon diversity of control group and metformin treated group, respectively, two sided t-test, $p = 0.84$) (Kemp and Aller, 2004). There are 46 significantly changed microbes, including 22 enriched microbes and 24 depleted microbes identified. At the class level, *Verrucomicrobiaceae*, *Prevotellaceae*, *Porphyromonadaceae*, *Rikenellaceae* are increased, while *Lachnospiraceae*, *Rhodobacteraceae* are decreased. Hierarchical clustering shows that samples from each group are clustered together (**Figure 1A**). Principal component analysis (PCA) suggests that metformin treated group and control group could be clearly separated in PC1, which explains 40.8% of the variation (**Figure 1B**). These results indicate that metformin consistently alters the gut microbiome of healthy mice.

To probe the function of these significantly changed microbes, PICRUSt was used to perform KEGG pathway analysis. Six pathways including ribosome, biosynthesis of amino acids, lipopolysaccharide biosynthesis, folate biosynthesis, purine metabolism and aminoacyl-tRNA biosynthesis are significant enriched (FDR < 0.05), see also **Figure 2**. The main function of gut microbiota is its important roles in metabolism, such as vitamin metabolism, short chain fatty acid metabolism, neuropeptide response, food digestion and so on. From the results of KEGG pathway analysis, it can be found that metformin mainly affects the gut microbiota related to such biological synthesis functions, including biosynthesis of lipopolysaccharide, folate, amino acids and proteins.

Comparative Analysis of the Altered Microbiome Profile

We applied our comparative microbiome tool MicroPattern to compare the significantly changed microbiome profile by metformin treatment, with the de-regulated microbiome profiles under various disease situations. Thirty-six and 29 significantly changed microbes of metformin treated T2DM patients (MTT2DM), from Forslund et al. and Wu et al.'s studies respectively, were integrated together for the analysis. We first calculated microbiome similarity between (MTT2DM) patients and other diseases. Then we calculated microbiome similarity between metformin treated healthy mice (MTHM) and other diseases. Finally, we also calculated microbiome similarity between prediabetes and other diseases. There are 17 diseases exhibit microbiome profile similarity with MTT2DM. Among them, 5 of them are significant (FDR < 0.1), see **Figure 3A**. As intuitively expected, we found that the metformin treated T2DM has the largest negative similarity (-0.26, FDR $<$ 1.00E-5) with T2DM. Metformin may reverse the

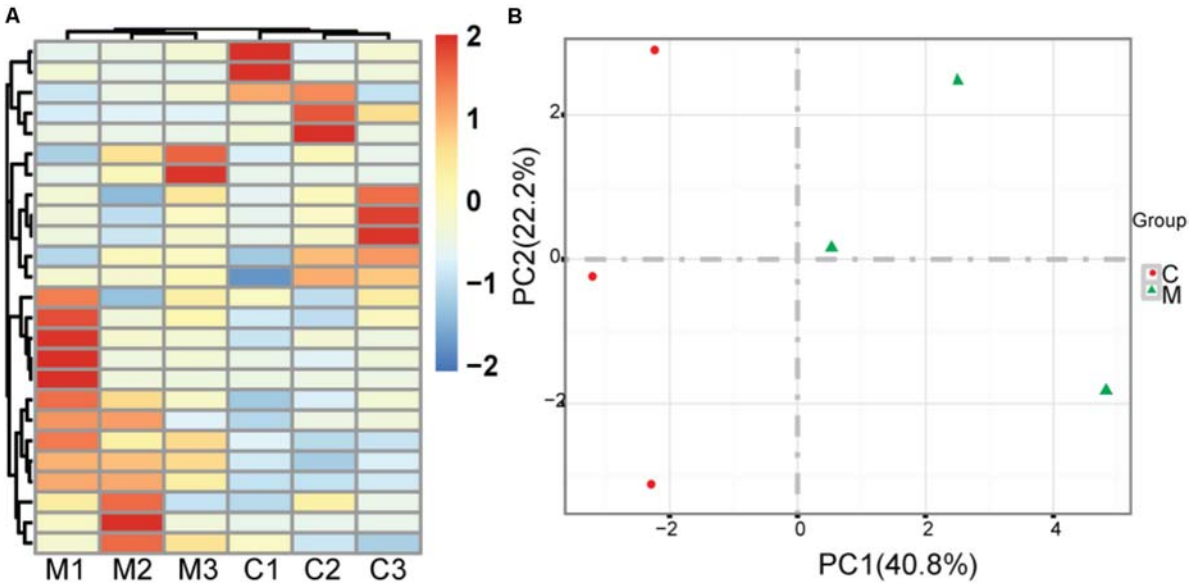


FIGURE 1 | Effect of metformin on the gut microbiome in healthy mice. **(A)** Heatmap of metformin treated mice and mock controls based on top 25 abundant microbes at the genus level; M: metformin treated mice; C: mock controls. **(B)** Principal component analysis of metformin treated mice and mock controls, red rounds indicate mock controls and green triangles indicate metformin treated mice; PC1: the first principal component; PC2: the second principal component.

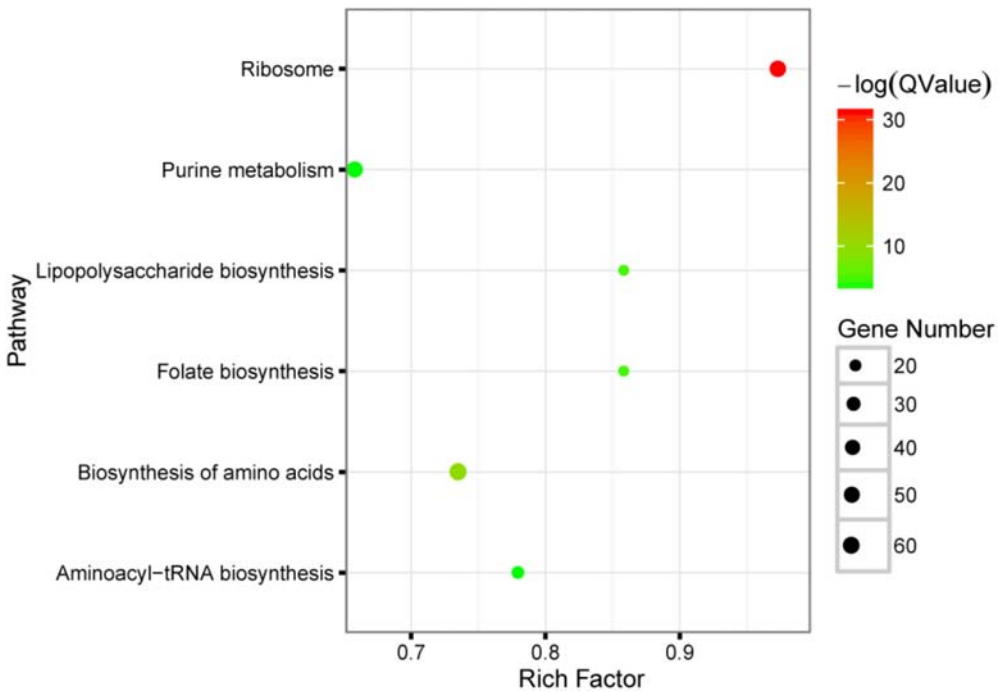


FIGURE 2 | KEGG pathway enrichment analysis of the significantly changed microbes with metformin treatment. The result shown in this bubble plot was generated by using PICRUSt tool. Briefly, each microbe was assigned to several specific metabolic gene signatures, and whether these genes were enriched in certain pathway was tested. The size of bubble plot is correlated with the number of the genes presented in the specific pathways while the color conforms to the statistical significance of the pathway enrichment; Rich factor: the number of significantly changed genes divided by the total number of genes in one pathway.

changed microbes under T2DM and thus relieves diabetic conditions. In contrast, there are 15 diseases that exhibit microbiome profile similarity with MTHM. Among them, 4 of them show positive similarities while 11 show negative similarities (**Figure 3B**). In all, five significant diseases are observed and 4 of them show negative similarity, indicating

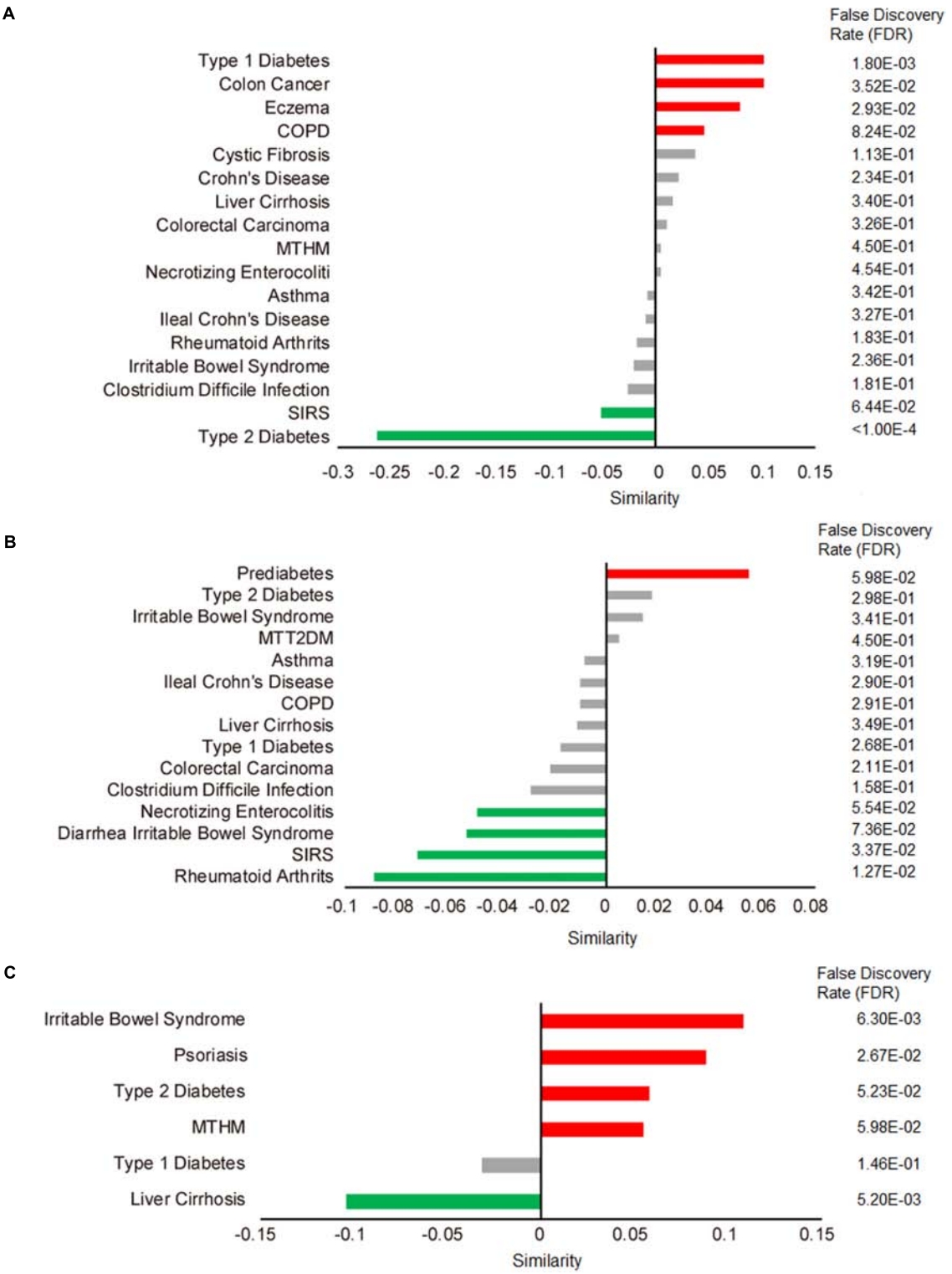


FIGURE 3 | Correlated microbiome profile changes between metformin treatment and disease conditions. Red bars indicate significant positive similarities (FDR < 0.1 by 10000 times of permutation-based resampling test), while green bar indicate significant negative similarities. Insignificant similarities are shown as gray bars. The FDR is listed in details on the right side. **(A)** Microbe profile similarity between metformin treated T2DM (MTT2DM) and other diseases. COPD, chronic obstructive pulmonary disease; SIRS, systemic inflammatory response syndrome. **(B)** Microbe profile similarity between metformin treated healthy mice (MTHM) and other disease. **(C)** Microbe profile similarity between prediabetes and other disease.

that metformin may play an important role in gut microbiota homeostasis by reversing the disease-associated microbiome alteration. Finally, as for prediabetes, there are 6 diseases exhibiting microbiome profile similarity (**Figure 3C**). In the original study, the authors investigated the aberrant of intestinal microbiota in individuals with prediabetes, overweight, insulin resistance, dyslipidaemia and low-grade inflammation, which were precursors of T2DM. Interestingly, the similarity between T2DM and prediabetes is 0.058, which suggests that the gut microbiota plays an important role in T2DM pre-conditioning. We also performed the enrichment analysis to identify the associated disease situations for the significantly changed microbes in MTT2DM, MTHM and prediabetes groups. No term is enriched in MTT2DM, rheumatoid arthritis and colorectal carcinoma are enriched in MTHM, whereas liver cirrhosis and irritable bowel syndrome are enriched in prediabetes (**Table 1**). From this result, MTHM show strong association with colorectal carcinoma, suggesting that the effect of anti-colorectal carcinoma of metformin may at least partly be mediated via gut microbiota. Moreover, the prediabetes individuals show altered microbiota profile similar to that in irritable bowel syndrome (**Figure 3B**), while MTHM negatively correlates with diarrhea irritable bowel syndrome in terms of microbiota alteration. Therefore, the metformin treatment may also be beneficial to the gut microbiota homeostasis as it can partly act against the microbiota de-regulation in prediabetes. Finally, MTHM also has negative similarity with multiple inflammatory diseases such as diarrhea irritable bowel syndrome, necrotizing enterocolitis, systemic inflammatory response syndrome (SIRS) and rheumatoid arthritis, which collectively indicates that the anti-inflammatory role of metformin is also related to gut microbiota.

DISCUSSION

Metformin was found useful for its anti-T2DM, anti-cancer, anti-aging effects and the treatment of polycystic ovary syndrome (Gundelach et al., 2016; Heckman-Stoddard et al., 2016; Kedia et al., 2016; Novelle et al., 2016). Previous researches have shown that gut microbiota alterations may be partly responsible for metformin's therapeutic effects against T2DM. For example, in diabetic rats, intravenous administration of metformin is less effective than intra-duodenal administration for lowering blood glucose levels (Bonora et al., 1984; Stepensky et al., 2002). Delayed-release metformin has lower bioavailability, and tends to accumulate in the lower bowel at higher concentrations compared with the common formulation (Stepensky et al., 2002; Buse et al., 2016). Changes of gut microbiota composition have been found in several diseases such as colon cancer (Wu et al., 2009; Kostic et al., 2013), rheumatoid arthritis (Scher et al., 2013), cardiovascular diseases (Wang et al., 2011; Tang et al., 2013) and diabetes (Larsen et al., 2010; Qin et al., 2012; Karlsson et al., 2013; Forslund et al., 2015; de la Cuesta-Zuluaga et al., 2017; Wu et al., 2017), obesity (Lee and Ko, 2014; Shin et al., 2014; Zhang et al., 2015). However, many questions exist. Could metformin alter gut microbiota of healthy individuals? How does metformin alter the gut microbiota of healthy individuals? What is the difference of the influence of metformin on gut microbiota under healthy and disease conditions? Does any correlation of microbiota alteration exist between different metformin treatment context? In our study, we treated healthy mice with metformin and found that metformin could indeed prominently affect gut microbiota under healthy condition. Subsequently, a computational method was applied for calculating the similarities between different conditions based on the changed microbes. Interestingly, the effects of metformin on gut microbiota turn out to be not identical under healthy and diabetes conditions. On the one hand, metformin could reverse the change of gut microbiota under diabetes, but the metformin treated mice did not show such trend. In fact, our result indicates a significant positive correlation with prediabetes and a weak positive correlation with T2DM. Therefore, although metformin shows a beneficial effect on gut microbiota in terms of improving disease condition in diabetic patients, our result cannot support the idea that metformin treatment of healthy mice could prevent diabetes-related gut microbiota disorder. To the contrary, metformin treatment of healthy mice may induce at least prediabetes. On the other hand, metformin treatment of diabetes patients positively correlates with colon cancer while metformin treatment of healthy mice exhibits negative correlation with multiple inflammatory diseases including diarrhea irritable bowel syndrome. This result indicates that metformin has potentially anti-inflammatory role especially under healthy condition. Indeed, the anti-inflammatory role of metformin was reported in previous research. For example, Koh et al. (2014) studied the anti-inflammatory mechanism of metformin but the proposed mechanism did not take gut microbiota into consideration. In their study, they found that metformin significantly inhibits interleukin (IL)-8 induction in COLO-205 cell stimulated with tumor necrosis factor (TNF)-α. Metformin significantly attenuates the severity of colitis in *IL-10*$^{-/-}$ mice and inhibits the development of colitic cancer in mice. Similarly, in Liu et al.'s study, metformin significantly decreases the mRNA expression of *IL-6* and *THF-α* and increases the mRNA expression of PI3K and Akt in pancreatic tissue of T2DM rats. A lot of microbes are shown to be correlated with inflammatory factors such as IL-6 and THF-α (Liu et al., 2018). In Lee et al. (2017) study, *IL-1*β and *IL-6* expression was significantly decreased in metformin-treated in aged obese mice and *IL-1*β and *IL-6* expression is negatively correlated with the abundance of *Bacteroides*, *Butyricimonas*, *Anaerotruncus* and *Akkermansia*. These studies link the anti-inflammatory mechanism of metformin with gut microbes in

TABLE 1 | MicroPattern disease enrichment analysis result.

Microbe set	*P*-value	FDR
MTHM		
Colorectal neoplasms	0.016	0.16
Arthritis, rheumatoid	0.016	0.16
Prediabetes		
Liver cirrhosis	0.0077	0.086
Irritable bowel syndrome	0.033	0.18

disease conditions. Our results also indicate that gut microbiota may play an important role in the anti-inflammatory effect of metformin in non-diabetic condition, which complements the conclusions of the previous studies. In summary, our microbiome profiling analysis signifies the role of gut microbiome in the mechanism underlying metformin treatment, which deserves detailed experimental and clinical investigation in the future.

DATA AVAILABILITY STATEMENT

The 16S rRNA sequencing datasets generated in this study can be found in the SRA database (https://www.ncbi.nlm.nih.gov/sra/?term=SRP099828).

AUTHOR CONTRIBUTIONS

WM performed the computational analysis. JC and YM performed the animal experiments. WM and YZ drafted the manuscript. QC, YZ, and JY conceived and designed the study. QC and YZ supervised the study.

FUNDING

This study was supported by the National High Technology Research and Development Program of China (Grant No. 2014AA021102 to QC), the National Natural Science Foundation of China (Grant Nos. 81422006 and 81670462 to QC), and China Postdoctoral Science Foundation (2016M591024 to YZ).

REFERENCES

Allin, K. H., Tremaroli, V., Caesar, R., Jensen, B. A. H., Damgaard, M. T. F., Bahl, M. I., et al. (2018). Aberrant intestinal microbiota in individuals with prediabetes. *Diabetologia* 61, 810–820. doi: 10.1007/s00125-018-4550-1

Bailey, C. J., and Turner, R. C. (1996). Metformin. *N. Engl. J. Med.* 334, 574–579. doi: 10.1056/nejm199602293340906

Banati, M., Csecsei, P., Koszegi, E., Nielsen, H. H., Suto, G., Bors, L., et al. (2013). Antibody response against gastrointestinal antigens in demyelinating diseases of the central nervous system. *Eur. J. Neurol.* 20, 1492–1495. doi: 10.1111/ene.12072

Bao, B., Wang, Z., Ali, S., Ahmad, A., Azmi, A. S., Sarkar, S. H., et al. (2012). Metformin inhibits cell proliferation, migration and invasion by attenuating CSC function mediated by deregulating miRNAs in pancreatic cancer cells. *Cancer Prev. Res.* 5, 355–364. doi: 10.1158/1940-6207.capr-11-0299

Berer, K., Mues, M., Koutrolos, M., Rasbi, Z. A., Boziki, M., Johner, C., et al. (2011). Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination. *Nature* 479, 538–541. doi: 10.1038/nature10554

Bolte, E. R. (1998). Autism and *Clostridium tetani*. *Med. Hypotheses* 51, 133–144. doi: 10.1016/S0306-9877(98)90107-4

Bonora, E., Cigolini, M., Bosello, O., Zancanaro, C., Capretti, L., Zavaroni, I., et al. (1984). Lack of effect of intravenous metformin on plasma concentrations of glucose, insulin, C-peptide, glucagon and growth hormone in non-diabetic subjects. *Curr. Med. Res. Opin.* 9, 47–51. doi: 10.1185/03007998409109558

Buse, J. B., DeFronzo, R. A., Rosenstock, J., Kim, T., Burns, C., Skare, S., et al. (2016). The primary glucose-lowering effect of metformin resides in the Gut, not the circulation: results from short-term pharmacokinetic and 12-week dose-ranging studies. *Diabetes Care* 39, 198–205. doi: 10.2337/dc15-0488

de la Cuesta-Zuluaga, J., Mueller, N. T., Corrales-Agudelo, V., Velasquez-Mejia, E. P., Carmona, J. A., Abad, J. M., et al. (2017). Metformin is associated with higher relative abundance of mucin-degrading *Akkermansia muciniphila* and several short-chain fatty acid-producing microbiota in the Gut. *Diabetes Care* 40, 54–62. doi: 10.2337/dc16-1324

Evans, J. M., Donnelly, L. A., Emslie-Smith, A. M., Alessi, D. R., and Morris, A. D. (2005). Metformin and reduced risk of cancer in diabetic patients. *BMJ* 330, 1304–1305. doi: 10.1136/bmj.38415.708634.F7

Ezendam, J., de Klerk, A., Gremmer, E. R., and van Loveren, H. (2008). Effects of *Bifidobacterium animalis* administered during lactation on allergic and autoimmune responses in rodents. *Clin. Exp. Immunol.* 154, 424–431. doi: 10.1111/j.1365-2249.2008.03788.x

Finegold, S. M., Dowd, S. E., Gontcharova, V., Liu, C., Henley, K. E., Wolcott, R. D., et al. (2010). Pyrosequencing study of fecal microflora of autistic and control children. *Anaerobe* 16, 444–453. doi: 10.1016/j.anaerobe.2010.06.008

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., et al. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 528, 262–266. doi: 10.1038/nature15766

Franciosi, M., Lucisano, G., Lapice, E., Strippoli, G. F., Pellegrini, F., and Nicolucci, A. (2013). Metformin therapy and risk of cancer in patients with type 2 diabetes: systematic review. *PLoS One* 8:e71583. doi: 10.1371/journal.pone.0071583

Furet, J. P., Kong, L. C., Tap, J., Poitou, C., Basdevant, A., Bouillot, J. L., et al. (2010). Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes Metab. Res. Rev.* 59, 3049–3057. doi: 10.2337/db10-0253

Garber, A. J., Abrahamson, M. J., Barzilay, J. I., Blonde, L., Bloomgarden, Z. T., Bush, M. A., et al. (2013). American Association of Clinical Endocrinologists' comprehensive diabetes management algorithm 2013 consensus statement–executive summary. *Endocr. Pract.* 19, 536–557. doi: 10.4158/ep13176.cs

Gundelach, T., Rodewald, M., Bekes, I., Janni, W., and Hancke, K. (2016). [Metformin for the treatment of polycystic ovary syndrome]. *Med. Monatsschr. Pharm.* 39, 75–78.

Heckman-Stoddard, B. M., Gandini, S., Puntoni, M., Dunn, B. K., DeCensi, A., and Szabo, E. (2016). Repurposing old drugs to chemoprevention: the case of metformin. *Semin. Oncol.* 43, 123–133. doi: 10.1053/j.seminoncol.2015.09.009

Hundal, R. S., Krssak, M., Dufour, S., Laurent, D., Lebon, V., Chandramouli, V., et al. (2000). Mechanism by which metformin reduces glucose production in type 2 diabetes. *Diabetes Metab. Res. Rev.* 49, 2063–2069. doi: 10.2337/diabetes.49.12.2063

Igel, L. I., Sinha, A., Saunders, K. H., Apovian, C. M., Vojta, D., and Aronne, L. J. (2016). Metformin: an old therapy that deserves a new indication for the treatment of obesity. *Curr. Atheroscler. Rep.* 18:16. doi: 10.1007/s11883-016-0568-3

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergstrom, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi: 10.1038/nature12198

Kedia, R., Kulkarni, S., Ross, M., and Shivaswamy, V. (2016). Spotlight on empagliflozin/metformin fixed-dose combination for the treatment of type 2 diabetes: a systematic review. *Patient Prefer. Adherence* 10, 1999–2006. doi: 10.2147/ppa.s85748

Kemp, P. F., and Aller, J. Y. (2004). Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiol. Ecol.* 47, 161–177. doi: 10.1016/s0168-6496(03)00257-5

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., et al. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* 19, 576–585. doi: 10.1038/nm.3145

Koh, S. J., Kim, J. M., Kim, I. K., Ko, S. H., and Kim, J. S. (2014). Anti-inflammatory mechanism of metformin and its effects in intestinal inflammation and colitis-associated colon cancer. *J. Gastroenterol. Hepatol.* 29, 502–510. doi: 10.1111/jgh.12435

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

Larsen, N., Vogensen, F. K., van den Berg, F. W., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., et al. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* 5:e9085. doi: 10.1371/journal.pone.0009085

Lee, H., and Ko, G. (2014). Effect of metformin on metabolic improvement and gut microbiota. *Appl. Environ. Microbiol.* 80, 5935–5943. doi: 10.1128/aem.01357-14

Lee, H., Lee, Y., Kim, J., An, J., Lee, S., Kong, H., et al. (2017). Modulation of the gut microbiota by metformin improves metabolic profiles in aged obese mice. *Gut Microbes* 9, 155–165. doi: 10.1080/19490976.2017.1405209

Lee, Y. K., Menezes, J. S., Umesaki, Y., and Mazmanian, S. K. (2011). Proinflammatory T-cell responses to gut microbiota promote experimental autoimmune encephalomyelitis. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4615–4622. doi: 10.1073/pnas.1000082107

Ley, R. E. (2010). Obesity and the human microbiome. *Curr. Opin. Gastroenterol.* 26, 5–11. doi: 10.1097/MOG.0b013e328333d751

Liu, G., Bei, J., Liang, L., Yu, G., Li, L., and Li, Q. (2018). Stachyose improves inflammation through modulating Gut microbiota of High-Fat Diet/Streptozotocin-induced Type 2 diabetes in rats. *Mol. Nutr. Food Res.* 62:e1700954. doi: 10.1002/mnfr.201700954

Lukens, J. R., Gurung, P., Vogel, P., Johnson, G. R., Carter, R. A., McGoldrick, D. J., et al. (2014). Dietary modulation of the microbiome affects autoinflammatory disease. *Nature* 516, 246–249. doi: 10.1038/nature13788

Ma, W., Huang, C., Zhou, Y., Li, J., and Cui, Q. (2017). MicroPattern: a web-based tool for microbe set enrichment analysis and disease similarity calculation based on a list of microbes. *Sci. Rep.* 7:40200. doi: 10.1038/srep40200

May, A., Brandt, B. W., El-Kebir, M., Klau, G. W., Zaura, E., Crielaard, W., et al. (2016). metaModules identifies key functional subnetworks in microbiome-related disease. *Bioinformatics* 32, 1678–1685. doi: 10.1093/bioinformatics/btv526

Novelle, M. G., Ali, A., Dieguez, C., Bernier, M., and de Cabo, R. (2016). Metformin: a hopeful promise in aging research. *Cold Spring Harb. Perspect. Med.* 6:a025932. doi: 10.1101/cshperspect.a025932

Ochoa-Reparaz, J., Mielcarz, D. W., Begum-Haque, S., and Kasper, L. H. (2011). Gut, bugs, and brain: role of commensal bacteria in the control of central nervous system disease. *Ann. Neurol.* 69, 240–247. doi: 10.1002/ana.22344

O'Hara, A. M., and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Rep.* 7, 688–693. doi: 10.1038/sj.embor.7400731

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. doi: 10.1038/nature11450

Restrepo, B. I. (2016). Metformin: candidate host-directed therapy for tuberculosis in diabetes and non-diabetes patients. *Tuberculosis* 101, S69–S72. doi: 10.1016/j.tube.2016.09.008

Scher, J. U., Sczesnak, A., Longman, R. S., Segata, N., Ubeda, C., Bielski, C., et al. (2013). Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* 2:e01202. doi: 10.7554/eLife.01202

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

Schwabe, R. F., and Jobin, C. (2013). The microbiome and cancer. *Nat. Rev. Cancer* 13, 800–812. doi: 10.1038/nrc3610

Shafiee, M. N., Khan, G., Ariffin, R., Abu, J., Chapman, C., Deen, S., et al. (2014). Preventing endometrial cancer risk in polycystic ovarian syndrome (PCOS) women: could metformin help? *Gynecol. Oncol.* 132, 248–253. doi: 10.1016/j.ygyno.2013.10.028

Shank, J. J., Yang, K., Ghannam, J., Cabrera, L., Johnston, C. J., Reynolds, R. K., et al. (2012). Metformin targets ovarian cancer stem cells in vitro and in vivo. *Gynecol. Oncol.* 127, 390–397. doi: 10.1016/j.ygyno.2012.07.115

Shin, N. R., Lee, J. C., Lee, H. Y., Kim, M. S., Whon, T. W., Lee, M. S., et al. (2014). An increase in the *Akkermansia* spp. population induced by metformin treatment improves glucose homeostasis in diet-induced obese mice. *Gut* 63, 727–735. doi: 10.1136/gutjnl-2012-303839

Sosnicki, S., Kapral, M., and Weglarz, L. (2016). Molecular targets of metformin antitumor action. *Pharmacol. Rep.* 68, 918–925. doi: 10.1016/j.pharep.2016.04.021

Stepensky, D., Friedman, M., Raz, I., and Hoffman, A. (2002). Pharmacokinetic-pharmacodynamic analysis of the glucose-lowering effect of metformin in diabetic rats reveals first-pass pharmacodynamic effect. *Drug Metab. Dispos.* 30, 861–868. doi: 10.1124/dmd.30.8.861

Tang, W. H., Wang, Z., Levison, B. S., Koeth, R. A., Britt, E. B., Fu, X., et al. (2013). Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* 368, 1575–1584. doi: 10.1056/NEJMoa1109400

Varrin-Doyer, M., Spencer, C. M., Schulze-Topphoff, U., Nelson, P. A., Stroud, R. M., Cree, B. A., et al. (2012). Aquaporin 4-specific T cells in neuromyelitis optica exhibit a Th17 bias and recognize *Clostridium* ABC transporter. *Ann. Neurol.* 72, 53–64. doi: 10.1002/ana.23651

Wang, Y., and Kasper, L. H. (2014). The role of microbiome in central nervous system disorders. *Brain Behav. Immun.* 38, 1–12. doi: 10.1016/j.bbi.2013.12.015

Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., Dugar, B., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* 472, 57–63. doi: 10.1038/nature09922

Wen, L., Ley, R. E., Volchkov, P. Y., Stranges, P. B., Avanesyan, L., Stonebraker, A. C., et al. (2008). Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* 455, 1109–1113. doi: 10.1038/nature07336

Witters, L. A. (2001). The blooming of the French lilac. *J. Clin. Invest.* 108, 1105–1107. doi: 10.1172/jci14178

Wu, H., Esteve, E., Tremaroli, V., Khan, M. T., Caesar, R., Manneras-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, 850–858. doi: 10.1038/nm.4345

Wu, L., Zhu, J., Prokop, L. J., and Murad, M. H. (2015). Pharmacologic therapy of diabetes and overall cancer risk and mortality: a meta-analysis of 265 studies. *Sci. Rep.* 5:10147. doi: 10.1038/srep10147

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

Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2365–2370. doi: 10.1073/pnas.0812600106

Zhang, H. H., and Guo, X. L. (2016). Combinational strategies of metformin and chemotherapy in cancers. *Cancer Chemother. Pharmacol.* 78, 13–26. doi: 10.1007/s00280-016-3037-3

Zhang, X., Zhao, Y., Xu, J., Xue, Z., Zhang, M., Pang, X., et al. (2015). Modulation of gut microbiota by berberine and metformin during the treatment of high-fat diet-induced obesity in rats. *Sci. Rep.* 5:14405. doi: 10.1038/srep14405

Zhang, Z. J., and Li, S. (2014). The prognostic value of metformin for cancer patients with concurrent diabetes: a systematic review and meta-analysis. *Diabetes Obes. Metab.* 16, 707–710. doi: 10.1111/dom.12267

Conflict of Interest Statement: 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.

Copyright © 2018 Ma, Chen, Meng, Yang, Cui and Zhou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 05 July 2018
doi: 10.3389/fmicb.2018.01411

Microbial Biogeography Along the Gastrointestinal Tract of a Red Panda

*Yan Zeng[1†], Dong Zeng[1†], Yi Zhou[1†], Lili Niu[2], Jiabo Deng[2], Yang Li[1], Yang Pu[2], Yicen Lin[1], Shuai Xu[1], Qian Liu[1], Lvchen Xiong[1], Mengjia Zhou[3], Kangcheng Pan[1], Bo Jing[1] and Xueqin Ni[1]**

[1] Animal Microecology Institute, College of Veterinary, Sichuan Agricultural University, Ya'an, China, [2] Chengdu Wildlife Institute, Chengdu Zoo, Chengdu, China, [3] Sichuan Animal Science Research Institute, Chengdu, China

The red panda (*Ailurus fulgens*) is a herbivorous carnivore that is protected worldwide. The gastrointestinal tract (GIT) microbial community has widely acknowledged its vital role in host health, especially in diet digestion; However, no study to date has revealed the GIT microbiota in the red panda. Here, we characterized the microbial biogeographical characteristics in the GIT of a red panda using high-throughput sequencing technology. Significant differences were observed among GIT segments by beta diversity of microbiota, which were divided into four distinct groups: the stomach, small intestine, large intestine, and feces. The stomach and duodenum showed less bacterial diversity, but contained higher bacterial abundance and the most unclassified tags. The number of species in the stomach and small intestine samples was higher than that of the large intestine and fecal samples. A total of 133 core operational taxonomic units were obtained from the GIT samples with 97% sequence identity. Proteobacteria (52.16%), Firmicutes (10.09%), and Bacteroidetes (7.90%) were the predominant phyla in the GIT of the red panda. Interestingly, *Escherichia–Shigella* were largely abundant in the stomach, small intestine, and feces whereas the abundance of *Bacteroides* in the large intestine was high. Overall, our study provides a deeper understanding of the gut biogeography of the red panda microbial population. Future research will be important to investigate the microbial culture, metagenomics and metabolism of red panda GIT, especially in *Escherichia–Shigella*.

Keywords: *Ailurus fulgens*, gastrointestinal tract, microbiota, *Escherichia–Shigella*, Illumina HiSeq sequencing

OPEN ACCESS

Edited by:
Xing Chen,
China University of Mining and Technology, China

Reviewed by:
Robert Heyer,
Universitätsklinikum Magdeburg, Germany
Young Min Kwon,
University of Arkansas, United States

****Correspondence:***
Xueqin Ni
xueqinni@foxmail.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: *10 February 2018*
Accepted: *08 June 2018*
Published: *05 July 2018*

Citation:
Zeng Y, Zeng D, Zhou Y, Niu L, Deng J, Li Y, Pu Y, Lin Y, Xu S, Liu Q, Xiong L, Zhou M, Pan K, Jing B and Ni X (2018) Microbial Biogeography Along the Gastrointestinal Tract of a Red Panda. Front. Microbiol. 9:1411. doi: 10.3389/fmicb.2018.01411

INTRODUCTION

The red panda (*Ailurus fulgens*) is a vulnerable wildlife species that belongs to the family *Ailuridae*, which is endemic to *Carnivora* (Yu et al., 2011). The species lives mainly in temperate forests in China, Bhutan, India, Burma, and Nepal. Their population is threatened by the climate, diet, and human activity (Deryabina et al., 2015; Princée and Glatston, 2016). A number of serious illnesses have led to death in this species, as investigated in the 20-year survey of captive dead red pandas (Delaski et al., 2015). Surveys show that pneumonia is the most common cause of death in newborns and juveniles, whereas cardiovascular disease, renal disease, and gastrointestinal disease are the common causes of death in the adult and geriatric red pandas. The probability of the survival of red panda has increased with successful captive breeding (Kumar et al., 2016). Improved captive measures, such as nutrition diets, regular veterinary care, and species breeding strategies, can promote the protection and conservation of red panda.

Intestinal microbial research is currently an important means of wildlife protection and conservation (Amato, 2013; Barelli et al., 2015; Stumpf et al., 2016). Microbiota plays a vital role in animal intestinal digestion, immune response, physiology, and disease treatment (Byndloss et al., 2017; Ramos and Hemann, 2017; Zheng et al., 2017). Together with the giant panda, the red panda is a herbivorous carnivore with simple gut morphologies (Ley et al., 2008). Nevertheless, they both specifically eat bamboo and shared 10 pseudogenes associated with digestion (Fei et al., 2017; Hu et al., 2017). Among the main candidate genes of the pseudothumbs, *DYNC2H1* and *PCNT* are mainly related to the absorption of amino acids in bamboo. Several studies have evaluated the faecal microbiota of red pandas and compared them with other wild animals, especially the giant panda (Kong et al., 2014; Li et al., 2015; Nishida and Ochman, 2017). Firmicutes was the predominant phylum found in the red and giant panda faecal, of which the bacteria abundance is extraordinary high in the giant panda. In particular, Proteobacteria was also found to be the second main flora in the red panda faecal. Firmicutes was found to be closely related to the degradation of bamboo fiber (Zhu et al., 2011). However, there is still relatively little research on the Proteobacteria in these animals gastrointestinal tract (GIT). No study has been conducted about the GIT microbiota of the red panda. All previous studies were conducted using faecal samples of red panda.

Our previous study assessed the bacterial diversity of this red panda using the polymerase chain reaction-denaturing gradient gel electrophoresis (DGGE) (Li et al., 2017). Higher bacterial diversity was found in the stomach and large intestine, whereas less bacterial diversity was obtained in the small intestine. Moreover, abundant DGGE bands in the GIT of red panda were identified with most belonging to Firmicutes, whereas the identified bacteria that belong to Proteobacteria were dominant in all segments. To improve our understanding of the microbial community structure and composition in red panda GIT, the Illumina HiSeq sequencing method is required. We hypothesize that the number and species of bacterial populations in the red panda's GIT is greater than what we know in it's faecal matter. We predict that Proteobacteria predominates in the GIT of the red panda. Our findings provide the first insight into the gut biogeography of microbial populations in red panda using high-throughput next-generation sequencing technology.

MATERIALS AND METHODS

Ethics Statement

The sample collection of the dead red panda was approved by the Ethical Committee of Animal Care and Use Commission of Chengdu Institute of Wildlife, Chengdu Zoo. Laboratory experiments were approved by the Animal Microecology Institute of Veterinary Medicine, Sichuan Agricultural University.

Sample Collection

In July 2016, a male, 5-year-old captive red panda was found dying in Chengdu Zoo. The animal was involved in a fight and its tail was docked before death. GIT contents were derived from GIT segments, which include the stomach, small intestine (duodenum, jejunum, and ileum), and large intestine (colon and rectum). The contents at the beginning and end of each segment were discarded. The content samples close to the middle of each segment were mixed. All GIT contents were collected within one day of the death of the red panda. In the same colonial house, faecal samples from other captive red pandas were thoroughly mixed to form a replacement faecal sample for the dead red panda. All samples were collected in accordance with the Sichuan Agricultural University Committee ethics (Certificate No. SYXKchuan 2014-187) regarding the care and use of experimental animals. The samples were placed in sterile tubes, frozen immediately, sent to the lab in 20 min, and stored at −80°C until further analysis.

DNA Extraction and Sequencing

Microbial genomic DNA was extracted from GIT contents and faecal samples (0.2 g each) using the QIAamp Stool Mini Kit (Qiagen, Germany). DNA concentration and purity were monitored on a Nano Drop spectrophotometer (Nano Drop Technologies, Wilmington, DE, United States) to ensure it is greater than 20 ng/μl and stored at −80°C prior to further analysis.

The V4 hypervariable region of the 16S rRNA gene from microbial genomic DNA was PCR-amplified using 515F (GTGCCAGCMGCCGCGGTAA) and 806R (GGAC TACHVGGGTWTCTAAT) primers with a 6 bp error-correcting barcodes (Caporaso et al., 2011) (**Supplementary Table S1**). PCR reactions were performed in triplicate with 20 μl of a mixture that contains 8 μl DNA, 1 μl each primer, 10 μl Phusion® High-Fidelity PCR Master Mix and 1 μl of ddH_2O. The following PCR reaction conditions were used: initial denaturation at 98°C for 1 min, 35 cycles of denaturation at 98°C for 10 s, annealing at 55°C for 30 s, and elongation at 72°C for 30 s and then 72°C for 5 min. PCR products were then mixed with the same volume of 1× loading buffer (contained SYBR green) and ran on a 2% agarose gel. PCR products with bright dominant bands of 400–450 bp were mixed at equal density ratios. The mixture PCR product was purified using a Qiagen gel extraction kit (Qiagen, Germany). Sequencing libraries were constructed according to the instructions using the TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, United States) and indexed by addition codes. Library quality was assessed using a Qubit@2.0 Fluorometer (Thermo Scientific) and Agilent Bioanalyzer 2100 system. Finally, sequencing was performed on the Illumina HiSeq 2500 platform, which generated 250 bp paired-end reads. The original 16S rRNA sequence data was available in the National Center for Biotechnology Information, BioProject ID PRJNA385220 and Sequencing Read Archive (SRP106218[1]).

Bioinformatics Analysis

The barcode and primer sequences (Caporaso et al., 2011) of paired-end sequencing readings in all samples were first

[1]https://www.ncbi.nlm.nih.gov/bioproject/PRJNA385220

removed. FLASH was used to assemble sample reads (V1.2.7[2]) (Magoč and Salzberg, 2011). Raw tags were filtered (Quality threshold < = 19, default length 3, and continuous high-quality base length greater than 75% tags) to obtain clean tags using QIIME (Version 1.7.0[3]) (Caporaso et al., 2010). To improve the quality of the analysis, a gold database[4] and UCHIME algorithm[5] (Edgar et al., 2011) were used to compare tags and remove chimeric sequences (**Supplementary Table S1**). Uparse software (Version 7.0.1001[6]) was used to cluster valid tags to obtain operational taxonomic units (OTUs) at 97% similarity level (Edgar, 2013). After removing the chloroplast and mitochondrial reads, species annotation (Threshold 0.8-1) was performed by the SSU rRNA database of SILVA[7] (Wang et al., 2007) and mothur (v 1.32) (Schloss et al., 2009). The classification level included the kingdom, phylum, class, order, family, genus, and species. Multiple sequence alignment analysis was performed using MUSCLE (Version 3.8.31[8]).

Alpha and beta diversity were analyzed using the QIIME (Version 1.7.0) (Caporaso et al., 2010) and visualized using R software (Version 2.15.3) (McMurdie and Holmes, 2013). With rarefaction at each sampling depth, alpha diversity included Shannon index, Simpson index, Observed-species, Good's coverage, Chao1, and ACE. Chao1 and ACE were obtained and show the community richness of the red panda GIT. The Shannon and Simpson indices revealed community diversity. The principal component analysis (PCA) of unweighted UniFrac distances was constructed (Lozupone et al., 2011). To compare the differences between the three groups, the ternary diagram was analyzed using the centroid plot of three variables, of which the sum of the three variables was constant (ggplot2) (Bulgarelli et al., 2015).

RESULTS

Metadata General Description

Using the Illumina HiSeq 2500 platform of 16S rRNA gene V4 region amplicons, a total of 460,679 sequences were obtained in seven samples of the red panda (**Supplementary Table S1**) with a median length of 253 bp. The number of sequences per sample ranged from 57,849 to 73,638. A total of 9,379 unique OTUs were obtained at 97% identity and an average of 1,340 OTUs for each sample, which range from 883 to 1643 (**Supplementary Figure S1A** and **Supplementary Table S1**). The number of OTUs was higher in the stomach (1643) and duodenum (1638). An average of 44 unclassified tags were observed in samples from the stomach (104 unclassified) and duodenum (121 unclassified) but not in the faecal samples. After the annotation of species through the SSU rRNA database, the taxonomic levels included kingdom, phylum, class, order, family, genus, and species of microbiota were conducted, which revealed in-depth microbial information (**Supplementary Figure S1B**). The large numbers of sequences in samples from the stomach and small intestine were detected at the species level. Samples from the large intestine revealed more sequences at the genus level. Species profiles observed from colon, rectum, and feces (**Supplementary Figure S2A**) tended to approach the saturated platform but also increased with samples from the stomach and small intestine. Arranging the rank abundance curve gave the same result and shows that only a large curve span is obtained from the colon, rectum, and feces (**Supplementary Figure S2B**). Generally, a large number of microbial sequences were investigated in the red panda, of which the number of the sequence is distinct among the GIT segments.

Microbial Diversity Across the Red Panda GIT

The alpha diversity (Observed species, Shannon, Simpson, Chao1, Ace, and goods coverage) was assessed by OTUs (**Supplementary Table S2**). Higher bacterial diversity was observed in both the large intestine and fecal samples than that in the stomach and small intestine. According to the ANOSIM analysis, significant differences were found in the bacterial community structure in the stomach, small intestine, large intestine, and faecal ($P < 0.05$). The number of species in the stomach and small intestine samples was higher than that of the large intestine and fecal samples. The beta diversity of the microbiota indicated that GIT segments are divided into four distinct groups: the stomach, small intestine, large intestine, and feces (**Figure 1A**). The heat map (weighted and unweighted uniFrac) of the distance matrix from the rectum and faecal samples showed a higher number than other samples (**Figure 1B**). This shows significant differences in the large intestine compared with other segments. A histogram generated by clustering analysis at phylum level divided the GIT samples into four major groups: the stomach, small intestine, large intestine, and feces (**Figure 1C**). Of the first three major bacterial phyla (Proteobacteria, Firmicutes, and Bacteroidetes), Proteobacteria predominated in all segments, especially in the stomach, small intestine and feces. Firmicutes was mainly distribute in faecal compared with other segments. Bacteroidetes was abundant in the large intestine.

Distinct Microbiota Across the Red Panda GIT

Next, the classification of specific taxonomy groups of species (e.g., kingdom, phylum, class, order, family, genus, and species) was conducted (**Figure 2** and **Supplementary Figure S3**). Proteobacteria (52.16%), Firmicutes (10.09%), and Bacteroidetes (7.90%) were the three major GIT phyla. *Escherichia–Shigella* (49.20%), *Helicobacter* (1.10%), *Pseudomonas* (1.07%), *Methylobacterium* (0.45%), and *Salinisphaera* (0.35%) mainly comprised the phylum Proteobacteria. Moreover, *Escherichia–Shigella* mainly included higher abundances of *Escherichia coli* (49.20%). *Escherichia–Shigella* showed the closest relationship with *Proteus* and *Morganella* in the

[2]http://ccb.jhu.edu/software/FLASH/
[3]http://qiime.org/scripts/split_libraries_fastq.html
[4]http://drive5.com/uchime/uchime_download.html
[5]http://www.drive5.com/usearch/manual/uchime_algo.html
[6]http://drive5.com/uparse/
[7]http://www.arb-silva.de/
[8]http://www.drive5.com/muscle/

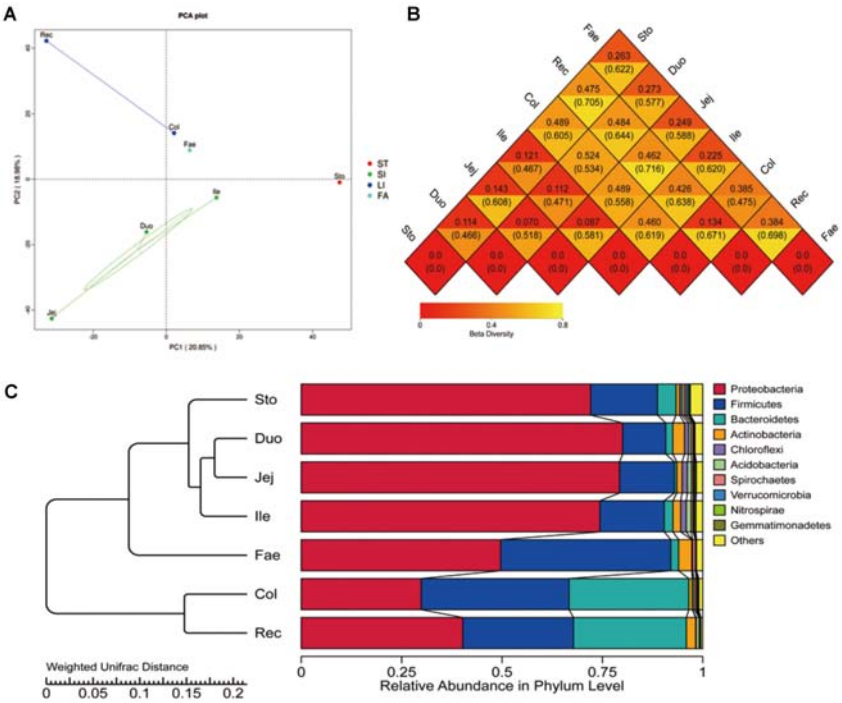


FIGURE 1 | Beta diversity of microbiota in red panda GIT. **(A)** Principal component analysis (PCA) with unweighted uniFrac of bacterial community in GIT samples. **(B)** Heatmap of the distance matrix, number above and under the central line in each rhombus is calculated using the weighted uniFrac and unweighted uniFrac methods, with higher values showing significant differences. **(C)** Clustering of the microbiota at the phylum level. Sto, Duo, Jej, Ile, Col, Rec, and Fae represent samples from the stomach, duodenum, jejunum, ileum, colon, rectum, and faecal, respectively.

evolutionary tree (**Supplementary Figure S4**). Firmicutes were primarily composed of *Enterococcus* (4.10%), *Clostridium_sensu_stricto_1* (3.02%), *Weissella* (2.62%), and *Turicibacter* (0.36%). The genus *Clostridium_sensu_stricto_1*, which mainly consisted *Clostridium_sp._CL-2* (0.83%), *Clostridium_beijerinckii* (0.14%), *Clostridium*_sp._ND2 (0.05%), *Clostridium_colicanis* (0.04%), and *Clostridium_perfringens* (0.03%), was the closest to the genus *sarcina* in the evolutionary tree (**Supplementary Figure S4**). Bacteroidetes was mainly composed of *Bacteroides* (7.89%), including *Bacteroides_fragilis* (3.67%), *Bacteroides_ovatus* (0.36%), *Bacteroides_uniformis* (0.30%), *Bacteroides_pyogenes* (0.05%), and *Bacteroides_caccae* (0.01%). Moreover, *Bacteroides* was the closest genus to *Alloprevotella*, *Prevotella_7*, and *Prevotella_9* in the phylogenetic tree (**Supplementary Figure S4**). The species annotation of red panda GIT microbiota was further analyzed at the family, genus, and species level using the SSU rRNA database (**Supplementary Figures S1B**, **S3**, **S5**). The composition of the entire red panda GIT, Enterobacteriaceae, Enterococcaceae, *Escherichia–Shigella*, *Enterococcus*, and *Escherichia coli* were enriched in the stomach and small intestine. Bacteroidaceae, Peptostreptococcaceae, Helicobacteraceae, Lachnospiraceae, Ruminococcaceae, Pseudomonadaceae, *Bacteroides*, *Helicobacter*, and *Pseudomonas* were mainly in the large intestines. Leuconostocaceae, *Weissella*, *Salinisphaera*, and *Turicibacter* were major in the faecal.

Using the Venn petal diagrams, a total of 133 core OTUs were obtained from the GIT samples of the red panda (**Figure 3A**). Of the 133 core OTUs, 100 genera of bacteria were identified. Of all the GIT samples, *Escherichia–Shigella* was the highest among the top 10 bacterial species, followed by *Bacteroides*, *Enterococcus*, *Clostridium_sensu_stricto_1*, *Helicobacter*, *Pseudomonas*, *Christensenellaceae* R-7 group, *Acinetobacter*, *Blautia*, and *Methylobacterium* (**Figure 3B** and **Supplementary Table S3**). The unique OTUs for stomach, duodenum, jejunum, ileum, colon, rectum, and faecal were 222, 158, 233, 158, 77, 179, and 86, respectively (**Figure 3A** and **Supplementary Table S4**). In the stomach sample, *Parafilimonas*, *Tamlana*, and *Thiocapsa* were the top three unique bacterial genera (**Figure 3C**). In the small intestine, *Aciditerrimonas*, *Inquilinus*, and unidentified_*Subgroup_7* were predominate in the duodenum, jejunum, and ileum, respectively (**Figures 3D–F**). *Polycyclovorans* and *Exiguobacterium* contribute most of the unique bacterial genus in colon and rectum samples. Moreover, *Nitrococcus*, *Filomicrobium*, and *Croceibacter* constituted the top three unique bacterial genera in faecal samples.

Predominant Bacteria With Classification From Phylum to Genus Revealed in the GIT of the Red Panda

Finally, the dominant bacteria belonged to the phylum Proteobacteria were analyzed in the red panda's GIT (**Figures 1**, **2**). The ternary plots of the bacteria of the stomach, small intestine, and large intestine samples at

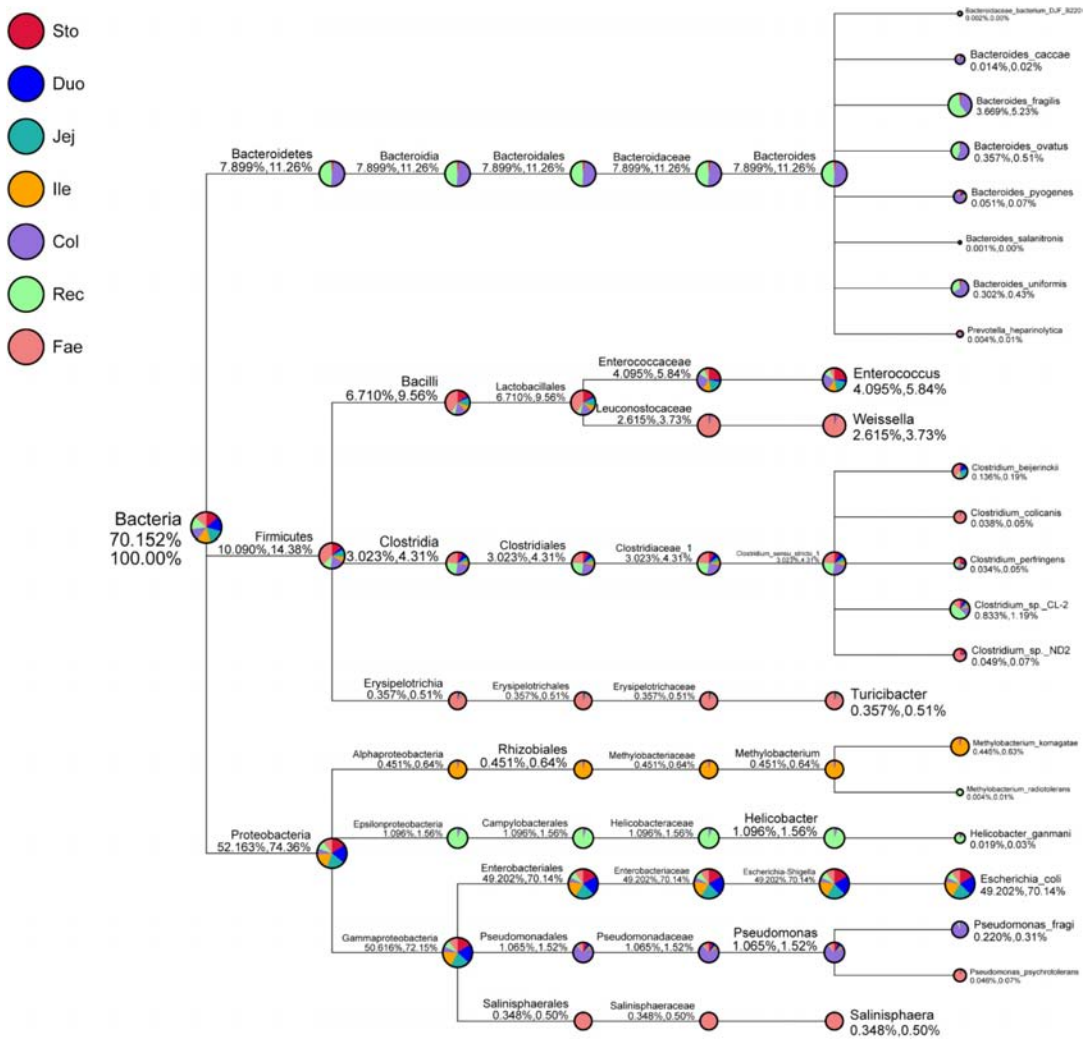


FIGURE 2 | Specific species taxonomy tree analysis of microbiota in red panda GIT. Classification levels include the kingdom, phylum, class, order, family, genus, and species. The first percentile in parentheses shows the percentage of microflora in all the detected bacteria. The second percentile in parentheses shows the percentage of microflora in all the selected bacteria. Sto, Duo, Jej, Ile, Col, Rec, and Fae represent samples from the stomach, duodenum, jejunum, ileum, colon, rectum, and faecal, respectively.

family and genus levels showed that the *Escherichia–Shigella* and Enterobacteriaceae were predominant and similar to the samples from the stomach and small intestine (**Figures 4A,B**). Bacteria sequenced due to the use of the 16S rRNA gene V4 region amplicons were less accurate. Thus, the statistical correlation of the *Escherichia–Shigella*, Enterobacteriaceae, Enterobacteriales, *Gammaproteobacteria*, and Proteobacteria across the GIT were exhibited in **Figure 4C**. These bacteria were present in the stomach and small intestine at a higher level than those in the large intestine and feces; the highest number was found in the duodenum. The duodenum bacterial sequence numbers for *Escherichia–Shigella*, Enterobacteriaceae, Enterobacteriales, *Gammaproteobacteria*, and Proteobacteria were 51,754, 52,527, 52,527, 53,525, and 56,810, respectively (**Supplementary Table S5**). We suggest that the bacteria are related to the small intestine, especially the duodenum.

DISCUSSION

Consistent with our hypothesis, the high-throughput sequencing data from the current study showed that the microbiota in the red panda GIT is distinct, with the Proteobacteria predominating (**Figure 3**). In healthy mammals, the stomach is the first segment of the GIT that receives and digests food and resides in bacteria that originated from Firmicutes, Bacteroidetes, and Proteobacteria (Gu et al., 2013; Gulino et al., 2013; Weldon et al., 2015). Moreover, the responsibility of nutrition digestion in small intestine is to ferment monosaccharides and amino acids (Gu et al., 2013). This gut nutrition environment is

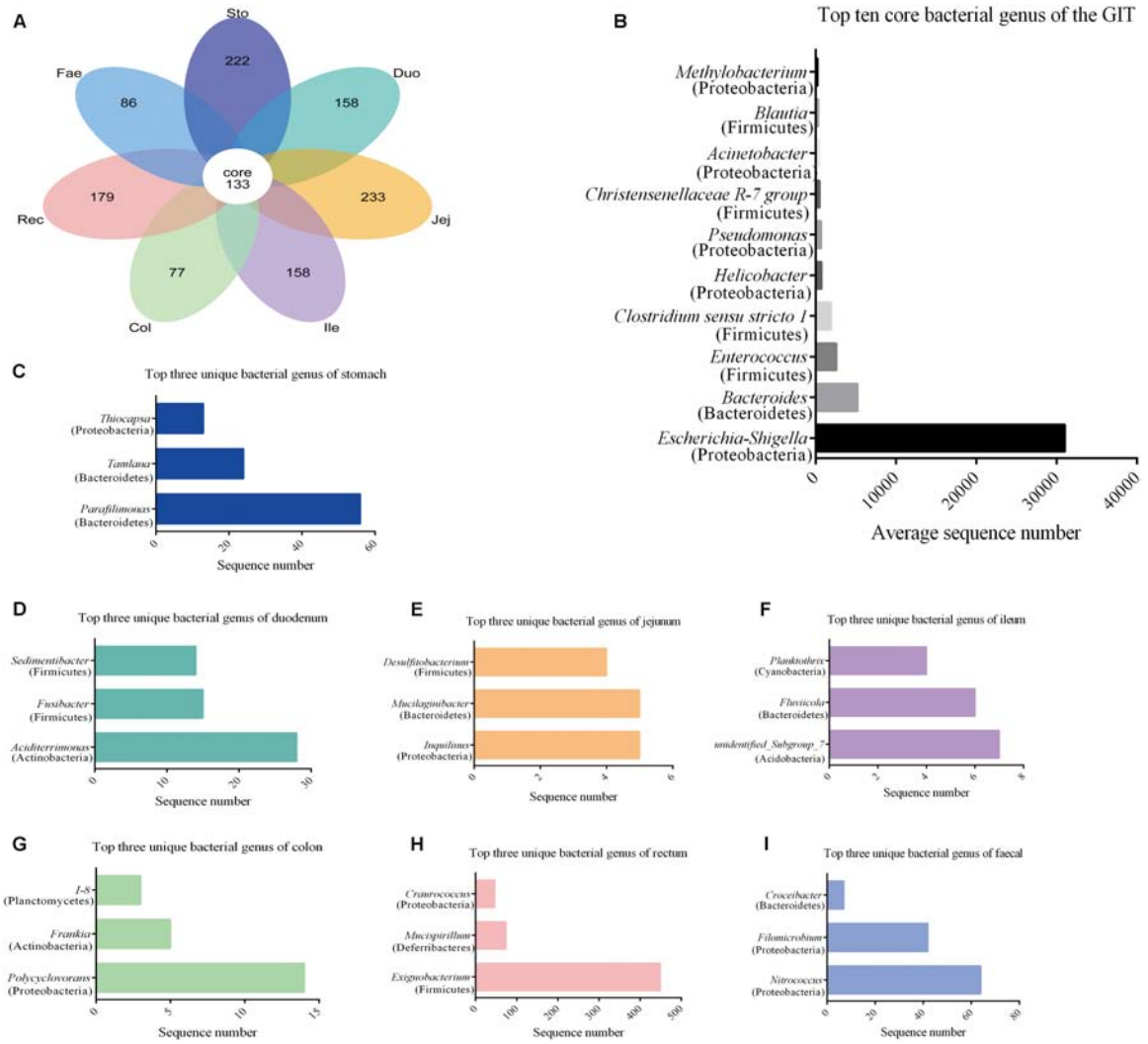


FIGURE 3 | Core and unique microbiota in red panda GIT. **(A)** The venn petal diagrams of the bacterial community in GIT. **(B)** Top 10 core bacterial genera of all the segments. **(C–I)** Top three unique bacterial genera in samples from the stomach, duodenum, jejunum, ileum, colon, rectum, and faecal, respectively.

suitable for facultative anaerobes growth, which mainly belongs to Proteobacteria. This existence and "disappearance" of the hypothetical "transient microbiota" may explain the great number of bacteria in the stomach and small intestine of mice, especially the highest numbers of duodenum bacteria (Gu et al., 2013). Our results revealed that the microbiota in the stomach and the small intestine of red panda showed a large number of OTUs: 1643 OTUs in the stomach and 1638 OTUs in the small intestine, mostly Proteobacteria (**Supplementary Figure S1A** and **Supplementary Table S1**). The large intestine digests polysaccharides and shows the dominance of Bacteroidetes (Faith et al., 2013; Seedorf et al., 2014). Bacteroidetes dominated the samples from the large intestine of the red panda, which is consistent with our findings. Although the number of bacterial sequences in these studies is not exactly the same, they have similar trends. Our data indicates that Firmicutes is the dominant species with a percentage of 40.49% in the fecal sample, compared with other segments. These results are consistent with the findings from other studies on microbiota in wild and captive red panda faecal (Kong et al., 2014; Li et al., 2015; Williams et al., 2018). In Williams's study, they used the Illumina MiSeq method to test the bacteria 16S rRNA V3–V4 region of two captive red panda faecal at different stages of weaning. Their results show that Firmicutes are the most abundant bacteria, with a percentage of 71 $\pm$ 6.9%. The Illumina MiSeq and Illumina HiSeq are two platforms widely used in the bacteria DNA sequencing, and they have 150 and 100 bp paired-end reads, respectively (Caporaso et al., 2012). The selection of the bacterial target regions of 16S rRNA can lead to the error and bias in amplicon-based microbial community (Gohl et al., 2016; Sinha et al., 2017). Nonetheless, these approaches allow the use of MiSeq and HiSeq methods to study the same trends of microbial flora.

For wild animals, bacterial communities in the feces are the easiest to study using living wild animals (Kong et al.,

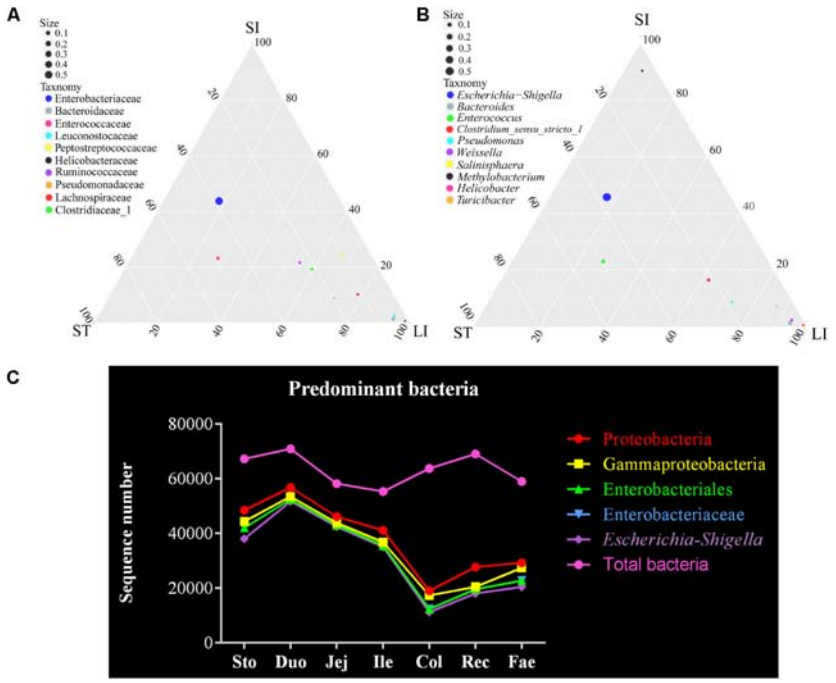


FIGURE 4 | *Escherichia–Shigella* is a major genus of bacteria in the red panda GIT. **(A)** The ternary plot of bacterial between the stomach, small intestine, and large intestine at family level. The three vertices in the figure represent the sample group (ST: stomach; SI: small intestinal; LI: large intestinal). Circles represent species, and size represents relative abundance. The closer the circle is to the vertex, the higher the relative abundance of this species. **(B)**The ternary plot of bacterial between the stomach, small intestine, and large intestine at genus level. **(C)** Total bacteria and the major bacterial genera of *Escherichia–Shigella*, classified from the phylum to genus level across the GIT.

2014; Barelli et al., 2015; Li et al., 2015; Borbón-García et al., 2017; McKenney et al., 2017; Menke et al., 2017). The study of GIT microbiota has broadened our understanding of host digestion, immune response, physiology and disease treatment and will help us further develope better ways of protecting wild animals (Amato, 2013; Bahrndorff et al., 2016; Stumpf et al., 2016). In recent years, some studies have focused on the study of the gastrointestinal flora of wild mammals. For example, when analyzing the V4 region of the 16S rRNA gene sequenced by Illumina MiSeq, the data showed that the colonized bacteria of macaques GIT are mainly Firmicutes, Bacteroidetes, Spirochaetes, and Proteobacteria (Yasuda et al., 2015). However, our study found that Proteobacteria is the major bacteria in the red panda GIT, followed by Firmicutes and Bacteroidetes (**Figure 1C**). Moreover, different results were found in the red kangaroo (*Macropus rufus*) GIT; Firmicutes, Bacteroidetes, and Actinobacteria were the major bacteria, and the target regions for 16S rRNA genes V3 and V4 were analyzed by Illumina MiSeq (Li et al., 2016). In addition to different sequencing methods, different breeds of animals contribute most of these differences. Similarly, Bacteroidetes (24.64%) and Firmicutes (13.28%) are mainly characterized in the GIT of a Brazilian Nelore steer. A similar result was found in the GIT of the bison (Bergmann, 2017), with the Bacteroidetes abundant in most segments. However, in the GIT of dairy cattle, the first three major bacteria were Firmicutes (42.22%), Bacteroidetes (21.00%), and Proteobacteria (17.56%). Bacterial relative abundance was found contributed most of the differences among different species of animals than that of the taxa of the bacteria (Chen et al., 2017). Firmicutes were found to be the dominant bacteria in the feces of primates (gibbon, golden monkey, chimpanzee, and assam macaque) and Proteobacteria were the dominant bacteria in carnivora (red panda, giant panda, tiger, black bear, and lion). Consistent with this study, our study found that Proteobacteria is primarily responsible for the red panda GIT, especially in the stomach and small intestine.

Our previous study showed that the predominate DGGE band in the GIT of the red panda was identified closest to the *Escherichia coli* strain KR-1. The same trend was found in this study. Proteobacteria was the main phyla in the GIT of red panda, including the class Gammaproteobacteria, the order Enterobacteriales, the family Enterobacteriaceae, and the genus *Escherichia–Shigella* (**Figure 4**). Microbiota appear to be functionally stable in the gut of different healthy hosts (Costea et al., 2017; Mehta et al., 2018). A recent study found that Proteobacteria (59.00%) is a dominant factor that influences functional variability in human gut microbiota than that in Bacteroidetes (12.00%) and Firmicutes (29.00%) (Bradley and Pollard, 2017). The class Gammaproteobacteria and Betaproteobacteria were found to be two of the four core microbiome in a survey of 112 animal species (herbivores, omnivores, and carnivores) representing 14 mammalian orders (Nishida and Ochman, 2017). However, Gammaproteobacteria

was observed more abundantly in captive animals in a study of the gut microbiome in 41 mammalian taxa (herbivores, omnivores, and carnivores) across six orders (Mckenzie et al., 2017). Despite this, Betaproteobacteria was found to be more abundant in wild animals in the six mammalian orders. This shift in Proteobacteria bacteria can be inferred to be related to animal diets, species and whether they are captive or wild. Moreover, Proteobacteria abundance seems to be sensitive to environmental changes. For example, Proteobacteria dominates in studies of fecal matter in brown bears (*Ursus arctos*) (Sommer et al., 2016) and Andean bears (Borbón-García et al., 2017). The change from winter to summer, frequent activities and food intake can lead to an increasing abundance of Proteobacteria in the feces of brown bears (*Ursus arctos*). Although captivity plays an important role in the population breeding and species conservation of wildlife, the bacterial diversity in their feces has declined (e.g., Andean bear, red panda, Przewalski's horse, woodrats, and panda) (Kohl et al., 2014; Kong et al., 2014; Wei et al., 2015; Borbón-García et al., 2017; Metcalf et al., 2017). These studies show that the protection of indigenous microbiota in the gut of wild animals is another important aspect of human conservation of wildlife. Currently, within its usefulness, *Escherichia–Shigella* can digest and absorb animal food. For example, *Escherichia–Shigella* is the dominant genus of feces in two captive red pandas during weaning (Williams et al., 2018). Moreover, *Escherichia–Shigella*, *Clostridium*, *Turicibacter*, and *Streptococcus* are the major genera in the wild giant panda feces at different times of the year, especially the utilization of the shoot and leaf stage (Wu et al., 2017). Consistent with this trend, *Escherichia–Shigella* is abundant in leaves that are predominantly mucus-seasoned samples (Williams et al., 2016). However, the overgrowth of Proteobacteria can also lead to some diseases, such as inflammatory bowel disease (Mukhopadhya et al., 2012) and metabolic syndrome (Shin et al., 2015). Additionally, *Escherichia–Shigella* is found to be closely related to *Proteus* and *Morganella* in the evolutionary tree in our study (**Figure 4C** and **Supplementary Figure S4**). A previous study shows that *Proteus* is associated with Crohn's disease (Lopetuso et al., 2018). With the distemper virus infection, the number of dominant *Escherichia* is reduced (Zhao et al., 2017). Based on the relatively few research results, we cannot confirm the role of the *Escherichia–Shigella* in red panda. However, recent studies have shown that different gastrointestinal (GI) diseases result in the significantly different composition of gut microbiota (Lopetuso et al., 2018). As for adult and old red pandas, gastrointestinal (e.g., ulceration, esophagitis, gastritis, diaphragmatic hernia, intussusception, and gastric torsion) and renal diseases (e.g., chronic interstitial nephritis and renal cysts) are mainly responsible for animal deaths (Delaski et al., 2015). Moreover, captive red pandas suffer from clinical illnesses, such as infectious diseases and parasites (Philippa and Ramsay, 2011). Gut microbes, such as *Clostridium*, *Lactobacillus*, *Eggerthella*, and *Bacteroides*, are usually active during the decomposing corpses (6–9 days) of dead bodies under natural conditions (Hyde et al., 2013). Although the GIT samples in our study are not fresh, they were collected within one day of the animal's death. Therefore, we consider that death status did not have significant impact on intestinal flora in the dead red panda in our study. Despite this, few intestinal samples of the dead red panda are available for further study, such as microbial culture and metabolomics studies. Not enough intestinal microflora information is available in other red panda studies for comparison. Given the similar habitat, dietary, and species evolution, the comparison of gut microbiota in other wildlife, such as the giant panda, is crucial to the GIT microbiota of red panda.

CONCLUSION

The contributions of this work are presented as follows: our study provides a first preliminary understanding of the biogeography of GIT microbiota in the red panda (*Ailurus fulgens*). Four different bacterial community areas, namely, the stomach, small intestine, large intestine, and feces, were obtained. Proteobacteria, Firmicutes, and Bacteroidetes were dominated by red panda's GIT. It will be important that future research investigate the microbial culture, metagenomics and metabolism of red panda GIT, especially in *Escherichia–Shigella*. Additionally, the results of the microbiota of the red panda in our study are limited. In the future, it will be necessary to conduct in-depth comparative analysis with other wild animals.

AUTHOR CONTRIBUTIONS

YaZ, XN, DZ, JD, LN, and YP conceived and designed the project. YiZ, YL, YcL, SX, and MZ for sample collection. YaZ, YL, QL, and LX performed experiments. YiZ, YL, KP, and BJ analysis and interpretation the data. YaZ, DZ, and XN wrote the manuscript. YaZ, YiZ, and DZ contributed to the manuscript equally. YiZ and XN revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by funding from the National Natural Science Foundation of China (31672318) and the Funded Project of Chengdu Giant Panda Breeding Research Foundation (CPF2017-04).

ACKNOWLEDGMENTS

The authors would like to thank the Chengdu Zoo veterinarian (Xingming Yu) for the anatomy of the dead animal and Chengdu Wildlife Institute for their excellent support. They also would like to thank Katherine Amato and Elizabeth Mallott at Northwestern University for the careful reading and revision of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.01411/full#supplementary-material

FIGURE S1 | Tag information and bacterial classification. **(A)** Total tags, taxon tags, operational taxonomic units (OTUs), unclassified tags, and unique tags. **(B)** The classification level includes the kingdom, phylum, class, order, family, genus, and species.

FIGURE S2 | Rarefaction curve analysis. Duplicate samples of OTU subgroups are evaluated whether further sampling would likely yield additional taxa, as indicated by the plateau value. **(A)** The y-axis represents the number of OTUs detected and the x-axis indicates the number of taxa in the analyzed subsets of sequences. **(B)** Rank abundance curves were used to estimate the richness and evenness of taxa present in the samples. The y-axis indicates the relative abundance of OTUs and the x-axis indicates the number of OTUs according to the relative abundance from largest to smallest. The larger the span curve is on the x-axis, the higher the species richness. The smoother the curve on the y-axis, the more evenly the species are distributed.

FIGURE S3 | Species-specific tree analysis of bacteria from the **(A)** stomach, **(B)** duodenum, **(C)** jejunum, **(D)** ileum, **(E)** colon, **(F)** rectum, and **(G)** faecal. The first percentile in brackets shows the percentage of all detected bacteria in the microbiota. The second percentile in parentheses shows the percentage of microbial communities in all selected bacteria.

FIGURE S4 | Top 100 bacteria genus in the evolutionary tree of red panda GIT.

FIGURE S5 | Species annotation of microbiota of red panda GIT at levels from **(A)** phylum, **(B)** class, **(C)** order, **(D)** family, and **(E)** genus.

TABLE S1 | General information of sequence data. Sto, Duo, Jej, Ile, Col, Rec, and Fae represent samples from the stomach, duodenum, jejunum, ileum, colon, rectum, and faecal, respectively.

TABLE S2 | Alpha diversity index, including observed species, Shannon, Simpson, chao1, ACE, and goods coverage. Sto, Duo, Jej, Ile, Col, Rec, and Fae represent samples from the stomach, duodenum, jejunum, ileum, colon, rectum, and faecal, respectively.

TABLE S3 | Core bacterial average sequence in red panda GIT. Sto, Duo, Jej, Ile, Col, Rec, and Fae represent samples from the stomach, duodenum, jejunum, ileum, colon, rectum, and faecal, respectively.

TABLE S4 | Unique bacterial genus from the stomach, duodenum, jejunum, ileum, colon, rectum, and faecal samples, respectively. Sto, Duo, Jej, Ile, Col, Rec, and Fae represent samples from the stomach, duodenum, jejunum, ileum, colon, rectum, and faecal, respectively.

TABLE S5 | Dominant bacterial community at the classification level from phylum to genus in red panda GIT.

REFERENCES

Amato, K. R. (2013). Co-evolution in context: the importance of studying gut microbiomes in wild animals. *Microb. Sci. Med.* 1, 10–29.

Bahrndorff, S., Alemu, T., Alemneh, T., and Lund Nielsen, J. (2016). The microbiome of animals: implications for conservation biology. *Int. J. Genomics* 2016:5304028. doi: 10.1155/2016/5304028

Barelli, C., Albanese, D., Donati, C., Pindo, M., Dallago, C., Rovero, F., et al. (2015). Habitat fragmentation is associated to gut microbiota diversity of an endangered primate: implications for conservation. *Sci. Rep.* 5:14862. doi: 10.1038/srep14862

Bergmann, G. T. (2017). Microbial community composition along the digestive tract in forage- and grain-fed bison. *BMC Vet. Res.* 13:253. doi: 10.1186/s12917-017-1161-x

Borbón-García, A., Reyes, A., Vivesflórez, M., and Caballero, S. (2017). Captivity shapes the gut microbiota of Andean bears: insights into health surveillance. *Front. Microbiol.* 8:1316. doi: 10.3389/fmicb.2017.01316

Bradley, P. H., and Pollard, K. S. (2017). Proteobacteria explain significant functional variability in the human gut microbiome. *Microbiome* 5, 36–58. doi: 10.1186/s40168-017-0244-z

Bulgarelli, D., Garrido-Oter, R., Münch, P. C., Weiman, A., Dröge, J., Pan, Y., et al. (2015). Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe* 17, 392–403. doi: 10.1016/j.chom.2015.01.011

Byndloss, M. X., Olsan, E. E., Rivera-Chávez, F., Tiffany, C. R., Cevallos, S. A., Lokken, K. L., et al. (2017). Microbiota-activated PPAR-γ signaling inhibits dysbiotic Enterobacteriaceae expansion. *Science* 357, 570–575. doi: 10.1126/science.aam9949

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, 335–336. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi: 10.1038/ismej.2012.8

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4516–4522. doi: 10.1073/pnas.1000080107

Chen, X., Li, Q. Y., Li, G. D., Xu, F. J., Han, L., Jiang, Y., et al. (2017). The distal gut bacterial community of some primates and carnivora. *Curr. Microbiol.* 75, 1–10. doi: 10.1007/s00284-017-1368-x

Costea, P. I., Hildebrand, F., Manimozhiyan, A., Bäckhed, F., Blaser, M. J., Bushman, F. D., et al. (2017). Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* 3, 8–16. doi: 10.1038/s41564-017-0072-8

Delaski, K. M., Ramsay, E., and Gamble, K. C. (2015). Retrospective analysis of mortality in the North American captive red panda (*Ailurus fulgens*) population, 1992–2012. *J. Zoo Wildl. Med.* 46, 779–788. doi: 10.1638/2014-0166.1

Deryabina, T. G., Kuchmel, S. V., Nagorskaya, L. L., Hinton, T. G., Beasley, J. C., Lerebours, A., et al. (2015). Long-term census data reveal abundant wildlife populations at Chernobyl. *Curr. Biol.* 25, 824–826.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

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

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., et al. (2013). The long-term stability of the human gut microbiota. *Science* 341:1237439. doi: 10.1126/science.1237439

Fei, Y., Hou, R., Spotila, J. R., Paladino, F. V., Qi, D., and Zhang, Z. (2017). Metabolic rate of the red panda, *Ailurus fulgens*, a dietary bamboo specialist. *PLoS One* 12:e0173274. doi: 10.1371/journal.pone.0173274

Gohl, D. M., Vangay, P., Garbe, J., Maclean, A., Hauge, A., Becker, A., et al. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* 34, 942–949. doi: 10.1038/nbt.3601

Gu, S., Chen, D., Zhang, J. N., Lv, X., Wang, K., Duan, L. P., et al. (2013). Bacterial community mapping of the mouse gastrointestinal tract. *PLoS One* 8:e74957. doi: 10.1371/journal.pone.0074957

Gulino, L. M., Ouwerkerk, D., Kang, A. Y., Maguire, A. J., Kienzle, M., and Klieve, A. V. (2013). Shedding light on the microbial community of the macropod foregut using 454-amplicon pyrosequencing. *PLoS One* 8:e61463. doi: 10.1371/journal.pone.0061463

Hu, Y., Wu, Q., Ma, S., Ma, T., Shan, L., Wang, X., et al. (2017). Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc. Natl. Acad. Sci. U.S.A.* 114, 1081–1086. doi: 10.1073/pnas.1613870114

Hyde, E. R., Haarmann, D. P., Lynne, A. M., Bucheli, S. R., and Petrosino, J. F. (2013). The living dead: bacterial community structure of a cadaver at the onset and end of the bloat stage of decomposition. *PLoS One* 8:e77733. doi: 10.1371/journal.pone.0077733

Kohl, K. D., Skopec, M. M., and Dearing, M. D. (2014). Captivity results in disparate loss of gut microbial diversity in closely related hosts. *Conserv. Physiol.* 2:cou009. doi: 10.1093/conphys/cou009

Kong, F., Zhao, J., Han, S., Zeng, B., Yang, J., Si, X., et al. (2014). Characterization of the gut microbiota in the red panda (*Ailurus fulgens*). *PLoS One* 9:e87885. doi: 10.1371/journal.pone.0087885

Kumar, A., Rai, U., Roka, B., Jha, A. K., and Reddy, P. A. (2016). Genetic assessment of captive red panda (*Ailurus fulgens*) population. *Springerplus* 5, 1750–1756.

Ley, R. E., Lozupone, C. A., Hamady, M., Hamady, M., Knight, R., and Gordon, J. I. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* 6, 776–788. doi: 10.1038/nrmicro1978

Li, M., Jin, W., Li, Y., Zhao, L., Cheng, Y., and Zhu, W. (2016). Spatial dynamics of the bacterial community structure in the gastrointestinal tract of red kangaroo (*Macropus rufus*). *World J. Microbiol. Biotechnol.* 32:98. doi: 10.1007/s11274-016-2054-z

Li, Y., Deng, J. B., Niu, L. L., Yu, J. Q., Zeng, Y., Liu, Q., et al. (2017). Bacteria diversity in gastrointestinal tract of *Ailurus fulgens* analyzed by polymerase chain reaction-denaturing gradient gel electrophoresis technology. *Chin. J. Anim. Nutr.* 29, 3167–3174.

Li, Y., Guo, W., Han, S., Kong, F., Wang, C., Li, D., et al. (2015). The evolution of the gut microbiota in the giant and the red pandas. *Sci. Rep.* 5:10185. doi: 10.1038/srep10185

Lopetuso, L. R., Petito, V., Graziani, C., Schiavoni, E., Paroni, S. F., Poscia, A., et al. (2018). Gut microbiota in health, diverticular disease, irritable bowel syndrome, and inflammatory bowel diseases: time for microbial marker of gastrointestinal disorders. *Dig. Dis.* 36, 56–65. doi: 10.1159/000477205

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172. doi: 10.1038/ismej.2010.133

Magoč, 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

McKenney, E. A., Maslanka, M., Rodrigo, A., and Yoder, A. D. (2017). Bamboo specialists from two mammalian orders (primates, carnivora) share a high number of low-abundance gut microbes. *Microb. Ecol.* 1–13. doi: 10.1007/s00248-017-1114-8

Mckenzie, V. J., Song, S. J., Delsuc, F., Prest, T. L., Oliverio, A. M., Korpita, T. M., et al. (2017). The effects of captivity on the mammalian gut microbiome. *Integr. Comp. Biol.* 57, 690–704. doi: 10.1093/icb/icx090

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217

Mehta, R. S., Abu-Ali, G. S., Drew, D. A., Lloyd-Price, J., Subramanian, A., and Lochhead, P. (2018). Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* 3, 347–355. doi: 10.1038/s41564-017-0096-0

Menke, S., Meier, M., Mfune, J. K. E., Melzheimer, J., Wachter, B., and Sommer, S. (2017). Effects of host traits and land-use changes on the gut microbiota of the Namibian black-backed jackal (*Canis mesomelas*). *FEMS Microbiol. Ecol.* 93:fix123. doi: 10.1093/femsec/fix123

Metcalf, J. L., Song, S. J., Morton, J. T., Weiss, S., Seguin-Orlando, A., Joly, F., et al. (2017). Evaluating the impact of domestication and captivity on the horse gut microbiome. *Sci. Rep.* 7:15497. doi: 10.1038/s41598-017-15375-9

Mukhopadhya, I., Hansen, R., El-Omar, E. M., and Hold, G. L. (2012). IBD-what role do Proteobacteria play? *Nat. Rev. Gastroenterol. Hepatol.* 9, 219–230. doi: 10.1038/nrgastro.2012.14

Nishida, A. H., and Ochman, H. (2017). Rates of gut microbiome divergence in mammals. *Mol. Ecol.* 12, 3218–3221. doi: 10.1111/mec.14473

Philippa, J., and Ramsay, E. (2011). "Chapter 15 – Captive red panda medicine," in *Red Panda*, ed. A. R. Glaston (London: Elsevier), 271–285. doi: 10.1016/B978-1-4377-7813-7.00015-X

Princée, F. P., and Glatston, A. R. (2016). Influence of climate on the survivorship of neonatal red pandas in captivity. *Zoo Biol.* 35, 321–330. doi: 10.1002/zoo.21266

Ramos, A., and Hemann, M. T. (2017). Drugs, bugs, and cancer: *Fusobacterium nucleatum* promotes chemoresistance in colorectal cancer. *Cell* 170, 411–413. doi: 10.1016/j.cell.2017.07.018

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

Seedorf, H., Griffin, N. W., Ridaura, V. K., Reyes, A., Cheng, J., Rey, F. E., et al. (2014). Bacteria from diverse habitats colonize and compete in the mouse gut. *Cell* 159, 253–266. doi: 10.1016/j.cell.2014.09.008

Shin, N. R., Whon, T. W., and Bae, J. W. (2015). Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends Biotechnol.* 33, 496–503. doi: 10.1016/j.tibtech.2015.06.011

Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., et al. (2017). Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* 35, 1077–1086. doi: 10.1038/nbt.3981

Sommer, F., Ståhlman, M., Ilkayeva, O., Arnemo, J. M., Kindberg, J., Josefsson, J., et al. (2016). The gut microbiota modulates energy metabolism in the hibernating brown bear *Ursus arctos*. *Cell Rep.* 14, 1655–1661. doi: 10.1016/j.celrep.2016.01.026

Stumpf, R. M., Gomez, A., Amato, K. R., Yeoman, C. J., Polk, J. D., Wilson, B. A., et al. (2016). Microbiomes, metagenomics, and primate conservation: new strategies, tools, and applications. *Biol. Conserv.* 199, 56–66.

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.

Wei, F., Wang, X., and Wu, Q. (2015). The giant panda gut microbiome. *Trends Microbiol.* 23, 450–452. doi: 10.1016/j.tim.2015.06.004

Weldon, L., Abolins, S., Lenzi, L., Bourne, C., Riley, E. M., and Viney, M. (2015). The gut microbiota of wild mice. *PLoS One* 10:e0134643. doi: 10.1371/journal.pone.0134643

Williams, C. L., Dill-McFarland, K. A., Sparks, D. L., Kouba, A. J., Willard, S. T., Suen, G., et al. (2018). Dietary changes during weaning shape the gut microbiota of red pandas (*Ailurus fulgens*). *Conserv. Physiol.* 6:cox075. doi: 10.1093/conphys/cox075

Williams, C. L., Dillmcfarland, K. A., Vandewege, M. W., Sparks, D. L., Willard, S. T., Kouba, A. J., et al. (2016). Dietary shifts may trigger dysbiosis and mucous stools in giant pandas (*Ailuropoda melanoleuca*). *Front. Microbiol.* 7:661. doi: 10.3389/fmicb.2016.00661

Wu, Q., Wang, X., Ding, Y., Hu, Y., Nie, Y., Wei, W., et al. (2017). Seasonal variation in nutrient utilization shapes gut microbiome structure and function in wild giant pandas. *Proc. R. Soc. B Biol. Sci.* 284:20170955. doi: 10.1098/rspb.2017.0955

Yasuda, K., Oh, K., Ren, B., Tickle, T. L., Franzosa, E. A., Wachtman, L. M., et al. (2015). Biogeography of the intestinal mucosal and lumenal microbiome in the rhesus macaque. *Cell Host Microbe* 17, 385–391. doi: 10.1016/j.chom.2015.01.015

Yu, L., Luan, P. T., Jin, W., Ryder, O. A., Chemnick, L. G., Davis, H. A., et al. (2011). Phylogenetic utility of nuclear introns in interfamilial relationships of Caniformia (order *Carnivora*). *Syst. Biol.* 60, 175–187. doi: 10.1093/sysbio/syq090

Zhao, N., Li, M., Luo, J., Wang, S., Liu, S., Wang, S., et al. (2017). Impacts of canine distemper virus infection on the giant panda population from the perspective of gut microbiota. *Sci. Rep.* 7:39954. doi: 10.1038/srep39954

Zheng, J. H., Nguyen, V. H., Jiang, S. N., Park, S. H., Tan, W., Hong, S. H., et al. (2017). Two-step enhanced cancer immunotherapy with engineered *Salmonella typhimurium* secreting heterologous flagellin. *Sci. Transl. Med.* 9:eaak9537. doi: 10.1126/scitranslmed.aak9537

Zhu, L., Wu, Q., Dai, J., Zhang, S., and Wei, F. (2011). Evidence of cellulose metabolism by the giant panda gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 108, 17714–17719. doi: 10.1073/pnas.1017956108

Conflict of Interest Statement: 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.

Copyright © 2018 Zeng, Zeng, Zhou, Niu, Deng, Li, Pu, Lin, Xu, Liu, Xiong, Zhou, Pan, Jing and Ni. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 12 September 2018
doi: 10.3389/fmicb.2018.02174

Sc-ncDNAPred: A Sequence-Based Predictor for Identifying Non-coding DNA in *Saccharomyces cerevisiae*

Wenying He[1], *Ying Ju*[2], *Xiangxiang Zeng*[2], *Xiangrong Liu*[2]* *and Quan Zou*[1,3]*

[1] School of Computer Science and Technology, Tianjin University, Tianjin, China, [2] School of Information Science and Technology, Xiamen University, Xiamen, China, [3] Shandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou, China

With the rapid development of high-speed sequencing technologies and the implementation of many whole genome sequencing project, research in the genomics is advancing from genome sequencing to genome synthesis. Synthetic biology technologies such as DNA-based molecular assemblies, genome editing technology, directional evolution technology and DNA storage technology, and other cutting-edge technologies emerge in succession. Especially the rapid growth and development of DNA assembly technology may greatly push forward the success of artificial life. Meanwhile, DNA assembly technology needs a large number of target sequences of known information as data support. Non-coding DNA (ncDNA) sequences occupy most of the organism genomes, thus accurate recognizing of them is necessary. Although experimental methods have been proposed to detect ncDNA sequences, they are expensive for performing genome wide detections. Thus, it is necessary to develop machine-learning methods for predicting non-coding DNA sequences. In this study, we collected the ncDNA benchmark dataset of *Saccharomyces cerevisiae* and reported a support vector machine-based predictor, called Sc-ncDNAPred, for predicting ncDNA sequences. The optimal feature extraction strategy was selected from a group included mononucleotide, dimer, trimer, tetramer, pentamer, and hexamer, using support vector machine learning method. Sc-ncDNAPred achieved an overall accuracy of 0.98. For the convenience of users, an online web-server has been built at: http://server.malab.cn/Sc_ncDNAPred/index.jsp.

Keywords: non-coding DNA, DNA sequence, feature representation, genome synthesis, support vector machine

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Chao Pang,
Columbia University Medical Center, United States
Qing Li,
University of Utah, United States

***Correspondence:**
Quan Zou
zouquan@tju.edu.cn
Xiangrong Liu
xrliu@xmu.edu.cn

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

Received: *24 July 2018*
Accepted: *24 August 2018*
Published: *12 September 2018*

Citation:
He W, Ju Y, Zeng X, Liu X and Zou Q (2018) Sc-ncDNAPred: A Sequence-Based Predictor for Identifying Non-coding DNA in Saccharomyces cerevisiae. Front. Microbiol. 9:2174. doi: 10.3389/fmicb.2018.02174

INTRODUCTION

After the implementation of many whole genome sequencing projects, more and more researches showed that non-coding DNA (ncDNA) is a major component of the biological genome. Numerous studies (Vogel, 1964; Thomas, 1971; Eddy, 2012; Puente et al., 2015; Liu et al., 2017a; Yao et al., 2018) have shown that the complexity of organisms is related to the length of non-coding regions, which are specially transcribed in physiological and disease states. Although the function of most ncDNAs is still unknown(Khurana et al., 2016), some studies (Horn et al., 2013; Huang et al., 2013; Vinagre et al., 2013; Puente et al., 2015; Hu et al., 2017, 2018; Rheinbay et al., 2017; Liao et al., 2018; Zhang W. et al., 2018) have shown that most cancer-related gene mutations are located in

ncDNA regions. How ncDNAs specifically affect tumor formation is also an urgent problem to be solved. In addition, ncDNAs in the genome play an important role in gene expressing, regulatory, and inheritance (Khurana et al., 2016).

Especially, with the rapid growth and development of synthetic biology, research in the genomics is advancing from genome sequencing to genome synthesis (Erlich and Zielinski, 2017; Jain et al., 2018; Liu B. et al., 2018). In recent years, various DNA assembly technologies (Ni et al., 2017; Wu et al., 2017; Xie et al., 2017; Zhang et al., 2017b) have been developed according to the principles of atypical enzyme cut connection (Engler et al., 2009; Sleight et al., 2010), single strand annealing and splicing (Gibson et al., 2009; Li and Elledge, 2012) and PCR (Warrens et al., 1997), which provide more rapid technical support for synthetic biology. In the following years, people are committed to improving the efficiency of large scale DNA assembly technologies. With the rapid development of the computer network and the popularity of the Internet, the number of digital information, such as network data, audio data, and video data, is increasing rapidly. It is urgent to establish a new system which has more efficiency than the existing storage system. DNA storage technology (Baum, 1995; Davis, 1996; Carr and Church, 2009) can meet the requirements above. In a new study (Shipman et al., 2017), the researchers introduced a method that encode images and video images into the genome of the *Escherichia coli* and read the corresponding images and videos from the genome of living bacterial cells. All the above studies require a large amount of DNA data.

As a complex type of genetic information, DNA sequences have specific characteristics not only in the coding sequence (cDNA) but also in the ncDNA sequences. Currently, the identification of cDNAs and ncDNAs relies mainly on experimental methods. However, traditional experimental methods are time-consuming and laborious, and the amount of genomic data is large and the sequence types are complex. In this context, there is an urgent need to establish accurate and efficient prediction methods to mine the information and knowledge of ncDNAs and cDNAs. Computational methods, which achieve a complementary effect, indeed effectively improved the recognition accuracy (Zhou et al., 2016).

In this study, a SVM-based computational method was first established to recognize the ncDNA sequences in *Saccharomyces cerevisiae (S. cerevisiae)*. Totally several types of features, such as mononucleotide composition (MNC), dimer nucleotide composition (DNC), trimer nucleotide composition (TNC), tetramer nucleotide composition (TrNC), pentamer nucleotide composition (PNC), and hexamer nucleotide composition (HNC) were extracted. The optimal feature extraction strategy was selected using SVM machine learning method. The workflow of constructing the Sc-ncDNAPred model is shown in **Figure 1**.

METHODS

Benchmark Dataset

In this study, the benchmark dataset was derived from the Ensembl genome database project (Hubbard et al., 2002), which is one of several well-known genome browsers for the retrieval of genomic information. Experimentally validated cDNA sequences of *S. cerevisiae* were extracted from their database, which contains 6713 samples. Intercepting the ncDNAs of the *S. cerevisiae* based on the initial marker information of the coding region provided by the original genomic data. By doing so, we obtained 6410 ncDNA samples. To get rid of redundancy, the CD-HIT (Li and Godzik, 2006) was adopted to remove those sequences that had $\geq$ 75% sequence identity. Finally, we obtained 6030 and 6251 samples in ncDNAs and cDNAs, respectively. Thus, the benchmark dataset can be formulated as

$$S = S^{+} \cup S^{-} \tag{1}$$

where S^{+} contained 6030 ncDNA samples, S^{-}contained 6251 cDNA samples and the symbol $\cup$ means the 'union' in the set theory.

The length distribution of ncDNA samples was shown in **Figure 2**. According to the graph, the length distribution of ncDNA is mainly between 100 and 800.

Feature Vector Construction

A sample can be simplified by a convenience form as:

$$\mathrm{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4 \ \ldots \ \mathrm{R}_{L-1}\mathrm{R}_L \tag{2}$$

where R_i ($i = 1,2,3 \ldots L$) represents the nucleotide at i-th position in one sequence.

K-mer Composition

K-mer nucleotide composition has been applied in many fields of bioinformatics (Liu et al., 2015b,c; Kim et al., 2017; Matias Rodrigues et al., 2017; Orenstein et al., 2017; Liu, 2018; Liu X. et al., 2018; Rangavittal et al., 2018). MNC equate to $k = 1$, DNC equate to $k = 2$, TNC equate to $k = 3$, TrNC equate to $k = 4$, PNC equate to $k = 5$, HNC equate to $k = 6$. The occurrence frequency of $k - mer(i)$can be represented as:

$$f_i^k = f(k - mer(i)) = \frac{n_i^k}{L - k + 1} \quad (i = 1, 2, ..., 4^k; k = 1, 2, 3, 4, 5, 6) \tag{3}$$

where n_i^kdenote the number of the i-th k-mer, L is the length of the sample sequence. Thus, each DNA sample can be defined feature vectors in different dimension of size 4^k. The generalized form of whole feature vectors X can be given by:

$$X = [f_1^k, f_2^k, \cdots, f_i^k, \cdots f_{4^k}^k]^T \tag{4}$$

Feature Ranking

Each sample sequence was represented by a large set of features, which leads to the redundant information (Wei and Billings, 2007; Senawi et al., 2017). In order to distinguish the contribution of different features to the prediction model. To analyze these feature vectors, *F-score* method (Chen W. et al., 2016; Jia and He, 2016; Tang et al., 2016, 2018; He and Jia, 2017) was adopted to

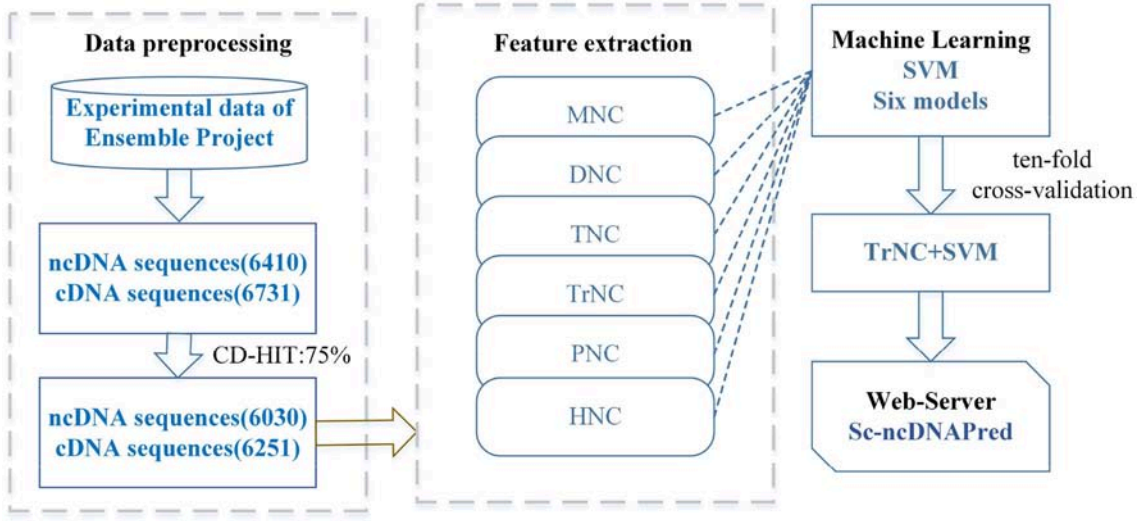


FIGURE 1 | The workflow of Sc-ncDNAPred.

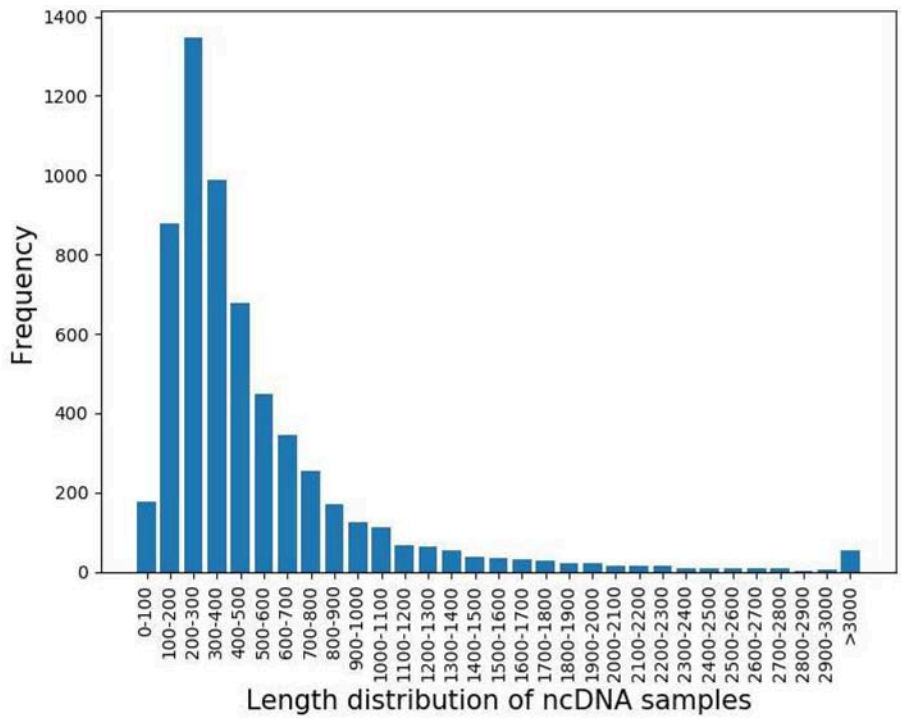


FIGURE 2 | The length distribution of ncDNA samples.

rank the feature, in this study. The *F-score* value of the i-th feature is defined as:

$$F{-}score(i) = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \bar{x}_i^{(-)} - \bar{x}_i{}^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n^--1}\sum_{k=1}^{n^-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (5)$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ are the average values of the i-th feature in whole, ncDNA and cDNA datasets, respectively. n^+ represents the number of ncDNA training samples, n^- represents the number of cDNA training samples, $x_{k,i}^{(+)}$ represents the i-th feature of the k-th ncDNA sample and $x_{k,i}^{(-)}$ represents the i-th feature of the k-th cDNA sample. Obviously, the feature with a greater score value indicates that it has a better discrimination ability.

Support Vector Machine

Support vector machine (SVM) (Hearst et al., 1998) is a widely used two-class classification algorithm based on statistical learning theory. It has been proven to be powerful in many fields of pattern recognition and data classification (Byun and Lee, 2002; Nasrabadi, 2007; Zhang N. et al., 2018;). More and more applications also proved that SVM also has strong data processing capabilities in the fields of bioinformatics (Xiong et al., 2011; Jia et al., 2013, 2017; Cao et al., 2014; Liu et al., 2014, 2017b; Wei et al., 2015; Chen X. X. et al., 2016; Jia and He, 2016; Yang et al., 2016; Zou et al., 2016; Xiao et al., 2017; Qiao et al., 2018; Su et al., 2018). A set of ncDNA samples and cDNA samples were represented by the feature vectors. The SVM classifies the data by mapping the input feature vectors to a high-dimensional feature space using a kernel function. In this study, the public LIBSVM package (Chang and Lin, 2011) was implemented to train models for discriminating between ncDNA sequences and cDNA sequences. Here, the radial basis function (RBF) $K(S_i, S_j) = exp(-\gamma \left\| S_i - S_j \right\|^2)$was set as the kernel function. The penalty parameter C and kernel parameter were preliminarily optimized through a grid search strategy.

Performance Evaluation

K-fold cross-validation (Chou and Zhang, 1995; Kohavi, 1995; Zhang et al., 2012a,b, 2015; Liu et al., 2015a; Chen X. et al., 2016; Li et al., 2016; Luo et al., 2016; Chen et al., 2017b, 2018a,b; Pan et al., 2017a; Xu et al., 2017; He et al., 2018) is one of the widely used approach to examine the ability of prediction model, and other approaches: independent dataset test and jackknife test (Chou and Shen, 2008) are also used in many applications. To reduce the computational cost, 10-fold cross validation was used to examine each model for its effectiveness in identifying ncDNA sequences. The training dataset were randomly divided into 10 subsets of approximately the same size. In each iteration, one subset was chosen as the test set and the remaining 9 subsets were used to train the model. For a complete cycle of a 10-fold cross-validation, the process was repeated 10 times until each subset was chosen as a test set. This 10-fold cross-validation procedure was repeated five times, then the results were averaged.

To evaluate the prediction performance of the models, five classic metrics were computed (Chou, 2001; Qiu et al., 2015, 2016; Liu et al., 2017; Pan et al., 2017b; Zhang et al., 2017a; Tang et al., 2018; Yang et al., 2018), including sensitivity (Sn), specificity (Sp), accuracy (Acc), Matthew correlation coefficient (MCC), and the receiver operating characteristic (ROC). These measurements were defined as:

$$
\begin{aligned}
Sn &= 1 - \frac{N_-^+}{N^+} \\
Sp &= 1 - \frac{N_+^-}{N^-} \\
Acc &= 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \\
MCC &= \frac{1 - (\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-})}{\sqrt{(1 + \frac{N_+^- - N_-^+}{N^+})(1 + \frac{N_-^+ - N_+^-}{N^-})}}
\end{aligned}
\tag{6}
$$

TABLE 1 | The 10-fold cross-validation results by different feature methods on the benchmark dataset.

Methods	Sn (%)	Sp (%)	ACC (%)	MCC
MNC	80.56	87.02	83.85	0.678
DNC	92.64	92.62	92.64	0.853
TNC	96.62	97.22	96.93	0.939
TrNC	98.01	98.51	98.26	0.965
PNC	95.25	95.84	95.56	0.911
HNC	90.71	92.25	91.49	0.830
All Features	95.99	96.08	96.03	0.921

The experiments have been executed 5 times and the results were the mean values.

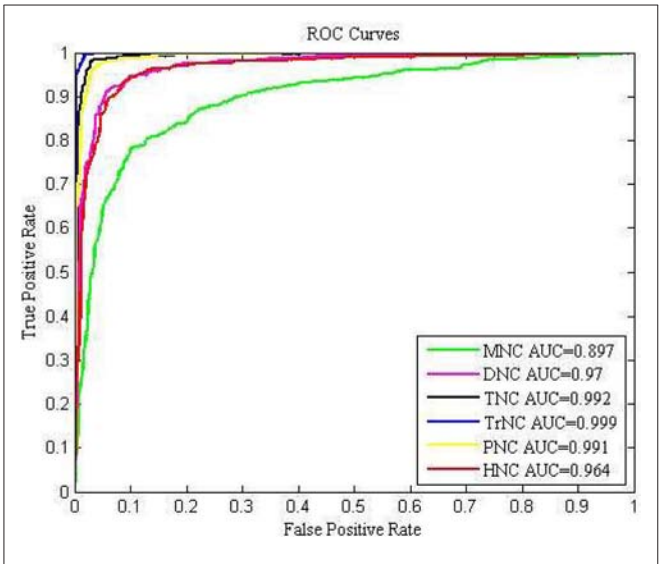


FIGURE 3 | The ROC curves to assess the predictive performance based on different feature extraction methods.

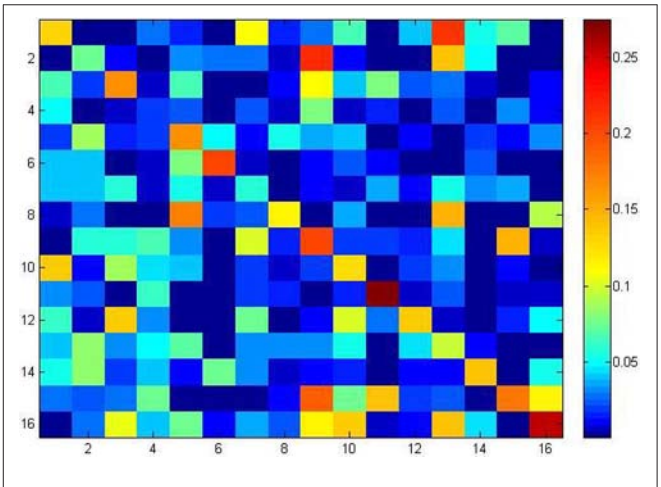


FIGURE 4 | Heap map to illustrate the *F_score* values of 256 different tetramers to identify ncDNA and cDNA.

TABLE 2 | Rules of composition of heat map.

AAAA	AAAC	AACA	AACC	ACAA	ACAC	ACCA	ACCC	CAAA	CAAC	CACA	CACC	CCAA	CCAC	CCCA	CCCC
AAAG	AAAT	AACG	AACT	ACAG	ACAT	ACCG	ACCT	CAAG	CAAT	CACG	CACT	CCAG	CCA	CCCG	CCCT
AAGA	AAGC	AATA	AATC	ACGA	ACGC	ACTA	ACTC	CAGA	CAGC	CATA	CATC	CCGA	CCGC	CCTA	CCTC
AAGG	AAGT	AATG	AATT	ACGG	ACGT	ACTG	ACTT	CAGG	CAG	CATG	CATT	CCGG	CCG	CCTG	CCTT
AGAA	AGAC	AGCA	AGCC	ATAA	ATAC	ATCA	ATCC	CGAA	CGAC	CGCA	CGCC	CTAA	CTAC	CTCA	CTCC
AGAG	AGAT	AGCG	AGCT	ATAG	ATAT	ATCG	ATCT	CGAG	CGAT	CGCG	CGCT	CTAG	CTAT	CTCG	CTCT
AGGA	AGGC	AGTA	AGTC	ATGA	ATGC	ATTA	ATTC	CGGA	CGGC	CGTA	CGTC	CTGA	CTGC	CTTA	CTTC
AGGG	AGGT	AGTG	AGTT	ATGG	ATGT	ATTG	ATTT	CGGG	CGGT	CGTG	CGTT	CTGG	CTGT	CTTG	CTTT
GAAA	GAAC	GACA	GACC	GCAA	GCAC	GCCA	GCCC	TAAA	TAAC	TACA	TACC	TCAA	TCAC	TCCA	TCCC
GAAG	GAAT	GACG	GACT	GCAG	GCAT	GCCG	GCCT	TAAG	TAAT	TACG	TACT	TCAG	TCAT	TCCG	TCCT
GAGA	GAGC	GATA	GATC	GCGA	GCGC	GCTA	GCTC	TAGA	TAGC	TATA	TATC	TCGA	TCGC	TCTA	TCTC
GAGG	GAGT	GATG	GATT	GCGG	GCGT	GCTG	GCTT	TAGG	TAGT	TATG	TATT	TCGG	TCGT	TCTG	TCTT
GGAA	GGAC	GGCA	GGCC	GTAA	GTAC	GTCA	GTCC	TGAA	TGAC	TGCA	TGCC	TTAA	TTAC	TTCA	TTCC
GGAG	GGAT	GGCG	GGCT	GTAG	GTAT	GTCG	GTCT	TGAG	TGAT	TGCG	TGCT	TTAG	TTAT	TTCG	TTCT
GGGA	GGGC	GGTA	GGTC	GTGA	GTGC	GTTA	GTTC	TGGA	TGGC	TGTA	TGTC	TTGA	TTGC	TTTA	TTTC
GGGG	GGGT	GGTG	GGTT	GTGG	GTGT	GTTG	GTTT	TGGG	TGGT	TGTG	TGTT	TTGG	TTGT	TTTG	TTTT

In these expressions, N^+ and N^- are the total number of ncDNA and cDNA samples, respectively, while N_-^+ and N_+^- are respectively the number of ncDNA samples incorrectly predicted as cDNA samples, and the number of cDNA samples incorrectly predicted as ncDNA samples.

RESULTS AND DISCUSSION

Prediction Results of Models

We used six types of effective feature extraction methods, such as MNC, DNA, TNC, TrNC, PNC, and HNC, as input of SVM to establish six models. The ability of each feature extraction method to discriminate between ncDNA and cDNA samples was compared by the 10-fold cross-validation (**Table 1**). As we can see from **Table 1**, the model for a combination SVM and TrNC yielded the best prediction performance, with the accuracy of 98.26%, the sensitivity of 98.01%, the specificity of 98.51%, and the MCC of 0.965, respectively. Then, the following second best prediction performance was yielded by TNC with the accuracy of 96.93%, the sensitivity of 96.62%, the specificity of 97.22%, and the MCC of 0.939, respectively. Besides, in the case of PNC, the corresponding model still obtained a good prediction results, which are 95.56% of accuracy, 95.25% of sensitivity, 95.84% of specificity and 0.911 of MCC, respectively.

To further investigate the overall prediction performance of each model, we showed the ROC curves and AUC values of different models for the 10-fold cross-validation in **Figure 3**. With the increase of *k*-mer value, the performance first increased and then decreased. Comparison demonstrated that the TrNC could produce the best results. Thus, the feature TrNC was adopted as the final model for Sc-ncDNAPred.

To further optimize the model, we performed multiple rounds of experiments on TrNC to select the appropriate subset of all 256 features (see Additional file 1: **Table S1** for full details); however, the results showed no significant improvement in the corresponding performance. The possible reason is that the selected feature cannot burden enough information for the discrimination.

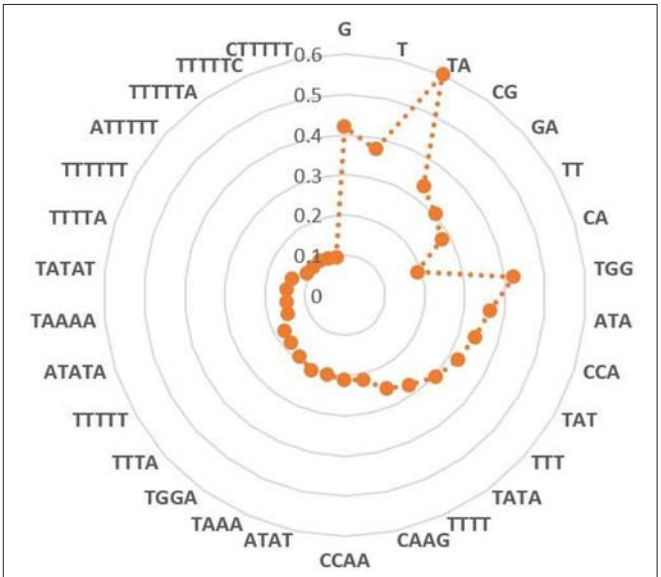


FIGURE 5 | Key features of each *k*-mer composition selected by *F-score* method. Red color denotes *F-score* value of each feature.

Compositional Analysis

To understand the 256 different tetramers bias in ncDNAs and cDNAs, a heap map was provided in **Figure 4**. Each square in the heat map corresponds to the *F-score* value of one tetramer (see **Table 2** for full details). Deep red in the heap map corresponds to a strong recognition ability.

Heap map analysis revealed that tetramers include TATA, TTTT, CAAG, CCAA, ATAT, TAAA, TGGA, TTTA, ATGG, ATAA, AATA, and CTGG are with the *F-score* values ranking

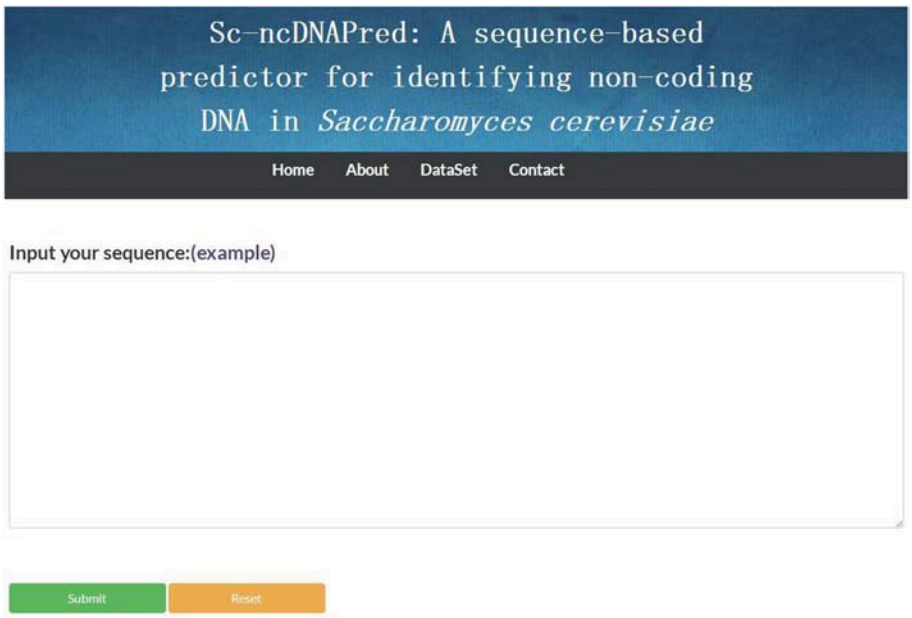


FIGURE 6 | A semi-screenshot of the top page of the Sc-ncDNAPred web-server at: http://server.malab.cn/Sc_ncDNAPred/index.jsp.

top twelve in all tetramers. In addition, we also analyzed the other *k*-mer components based on the *F-score* method, respectively. Among them, the two key nucleotides G and T from MNC, the top five key dimer nucleotide composition (TA, CG, GA, TT, and CA) from DNC, (TGG, ATA, CCA, TAT, and TTT) from TNC, (TTTTT, ATATA, TAAAA, TATAT, and TTTTA) from PNC, and (TTTTTT, ATTTTT, TTTTTA, TTTTTC and CTTTTT) from HNC. These key features are presented in a radar diagram (**Figure 5**). The study of these key features can deepen the understanding of the overall structure of the genome, which not only promotes the annotation of the genome, but also promotes the study of biological evolution.

Comparison With Other Classifiers

To the best of our knowledge, this is the first time that machine learning method has been used to identify ncDNA in S. cerevisiae. In order to further testify the superiority of proposed model Sc-ncDNAPred, the predictive results of it were compared with that of other powerful and widely used classifiers, i.e., k-Nearest Neighbor (KNN), Naïve Bayes, Random Forest, and J48 Tree as implemented in WEKA (Frank et al., 2004).The 10-fold cross validation results of these four classifier for identifying ncDNA in the same benchmark dataset were shown in Additional file 1: **Table S2**. The results showed that the four metrics as defined in Eq. 6 of the proposed model Sc-ncDNAPred are all higher than those of k-Nearest Neighbor (KNN), Naïve Bayes, Random Forest, and J48 Tree.

Web-Server

Based on the benchmark dataset defined in Eq.1, a predictor called Sc-ncDNAPred was established, where "Sc" stands for *S. cerevisiae* and "Pred" stands for "Prediction." For conveniences of users' community, a step-by-step guide about how to use the web-server is provided as follows:

Step 1. Open the web-server at: http://server.malab.cn/Sc_ncDNAPred/index.jsp, you will see the home page of Sc-ncDNAPred, as shown in **Figure 6**. Click the "About" button to see a brief introduction of the server.
Step 2. Paste the query DNA sequences into the input box. The input sequence should be in FASTA format. For the example of DNA sequences in FASTA format, click the "example" button top above the input box.
Step 3. Click on the "Submit" button to start the prediction. If the prediction result of a sequence is positive, its output is "ncDNA." Otherwise, its output is "cDNA."
Step 4. Click on the "DataSet" button to download the benchmark dataset.
Step 5. Click on the "Contact" button to contact us.

CONCLUSIONS

DNA assembly technology needs a large number of target sequences of known information as data support. Non-coding DNA (ncDNA) sequences occupy most of the organism genomes, thus accurate recognizing of them is necessary. In this study, an efficient computational model was proposed to identify ncDNAs in *S. cerevisiae*. The tetramer nucleotide composition (TrNC) was adopted to extract features. The *F-score* method was used to analyze these feature vectors and find the key features. The high accuracy indicated that Sc-ncDNAPred was a powerful tool for predicting ncDNA. Finally, a free web-server was developed based on the proposed model. We hope

that the predictor will provide convenience to most of scholars. Currently, annotations for the genomic sequences of most species are lacking or unavailable. To analyze the ncDNA data of these organisms, we can obtain data and methodological support in a cross-species manner from annotated species. For example, we could try to use the model built from *S. cerevisiae* dataset to analyze other species of bacteria that have not been explored in depth. In addition, we will also apply this computational model for the prediction of potential disease related non-coding DNA. In the future, we will apply this computational model for the prediction of potential disease related non-coding RNA (Chen and Huang, 2017; Chen et al., 2017a, 2018c,d; You et al., 2017).

AUTHOR CONTRIBUTIONS

WH, QZ, and XL wrote the paper. XZ and YJ participated in preparation of the manuscript. QZ, WH, XL, XZ, and YJ participated in the research design. WH and QZ developed the web server. WH, YJ, XZ, XL, and QZ read and approved the final manuscript.

FUNDING

The work was supported by the National Natural Science Foundation of China (Nos. 61771331, 61472333, 61772441, 61472335, 61425002), Funding from Shandong Provincial Key Laboratory of Biophysics, Project of marine economic innovation and development in Xiamen (No. 16PFW034SF02), Natural Science Foundation of the Higher Education Institutions of Fujian Province (No. JZ160400), Natural Science Foundation of Fujian Province (No. 2017J01099), President Fund of Xiamen University (No. 20720170054), and Shenzhen Overseas High Level Talents Innovation Foundation (No. KQJSCX20170327161949608). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02174/full#supplementary-material

REFERENCES

Baum, E. B. (1995). Building an associative memory vastly larger than the brain. *Science* 268, 583–585.

Byun, H., and Lee, S. W. (2002). Applications of support vector machines for pattern recognition: a survey. In: *Pattern Recognition With Support Vector Machines. Springer p* (Niagara Falls, ON), 213–236.

Cao, R., Wang, Z., Wang, Y. H., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinform.* 15:120. doi: 10.1186/1471-2105-15-120

Carr, P. A., and Church, G. M. (2009). Genome engineering. *Nat. Biotechnol.* 27, 1151–1162. doi: 10.1038/nbt.1590

Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K. C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 7, 16895–16909. doi: 10.18632/oncotarget.7815

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018a). EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.* 9:3. doi: 10.1038/s41419-017-0003-x

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2018b). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 34:1440. doi: 10.1093/bioinformatics/btx773

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018c). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics*. doi: 10.1093/bioinformatics/bty503. [Epub ahead of print].

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018d). BNPMDA: bipartite network projection for MiRNA-disease association prediction. *Bioinformatics*. doi: 10.1093/bioinformatics/bty333. [Epub ahead of print].

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2017a). MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* doi: 10.1093/bib/bbx130. [Epub ahead of print].

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017b). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016). Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed. Res. Int.* 2016:1654623. doi: 10.1155/2016/1654623

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Bioinform.* 43, 246–55. doi: 10.1002/prot.1035

Chou, K. C., and Shen, H. B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. doi: 10.1038/nprot.2007.494

Chou, K. C., and Zhang, C. T. (1995). Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* 30, 275–349.

Davis, J. (1996). Microvenus. *Art J.* 55, 70–74.

Eddy, S. R. (2012). The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* 22, R898–R899. doi: 10.1016/j.cub.2012.10.002

Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009). Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PloS ONE* 4:e5553. doi: 10.1371/journal.pone.0005553

Erlich, Y., and Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. *Science* 355, 950–954. doi: 10.1126/science.aaj2038

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481. doi: 10.1093/bioinformatics/bth261

Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A. III., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6, 343–345. doi: 10.1038/nmeth.1318

He, W., and Jia, C. (2017). EnhancerPred2. 0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron–ion interaction potential feature selection. *Mol BioSyst.* 13, 767–74. doi: 10.1039/c7mb00054e

He, W., Jia, C., Duan, Y., and Zou, Q. (2018). 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol.* 12:44. doi: 10.1186/s12918-018-0570-1

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intell. Syst. Appl.* 13, 18–28.

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science* 339, 959–961. doi: 10.1126/science.1230062

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* doi: 10.1080/15476286.15472018.11457935. [Epub ahead of print].

Hu, H., Zhu, C., Ai, H., Zhang, L., Zhao, J., Zhao, Q., et al. (2017). LPI-ETSLP: lncRNA-protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction. *Mol. Biosyst.* 13, 1781–1787. doi: 10.1039/c7mb00290d

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., and Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science* 339, 957–959. doi: 10.1126/science.1229259

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., et al. (2002). The ensembl genome database project. *Nucleic Acids Res.* 30, 38–41. doi: 10.1093/nar/30.1.38

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. doi: 10.1038/nbt.4060

Jia, C., and He, W. (2016). EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. *Sci. Rep.* 6:38741. doi: 10.1038/srep38741

Jia, C. Z., He, W. Y., and Yao, Y. H. (2017). OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids. *J. Biomol. Struct. Dyn.* 35, 829–835. doi: 10.1080/07391102.2016.1163294

Jia, C. Z., Liu, T., and Wang, Z. P. (2013). O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. *Mol. Biosyst.* 9, 2909–2913. doi: 10.1039/C3MB70326F

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* 17, 93–108. doi: 10.1038/nrg.2015.17

Kim, C. S., Winn, M. D., Sachdeva, V., and Jordan, K. E. (2017). K-mer clustering algorithm using a MapReduce framework: application to the parallelization of the Inchworm module of Trinity. *BMC Bioinform.* 18:467. doi: 10.1186/s12859-017-1881-8

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. in *Ijcai 95 Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, QC.* 1137–1145.

Li, D., Luo, L., Zhang, W., Liu, F., and Luo, F. (2016). A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinform.* 17:329. doi: 10.1186/s12859-016-1206-3

Li, M. Z., and Elledge, S. J. (2012). SLIC: a method for sequence- and ligation-independent cloning. *Methods Mol. Biol.* 852, 51–59. doi: 10.1007/978-1-61779-564-0_5

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

Liao, Z. J., Li, D. P., Wang, X. R., Li, L. S., and Zou, Q. (2018). Cancer diagnosis through IsomiR expression with machine learning method. *Curr. Bioinform.* 13, 57–63. doi: 10.2174/1574893611666160609081155

Liu, B. (2018). BioSeq-analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* doi: 10.1093/bib/bbx165. [Epub ahead of print].

Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., and Chou, K. C. (2015a). Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS ONE* 10:e0121501. doi: 10.1371/journal.pone.0121501

Liu, B., Fang, Y., Huang, D. S., and Chou, K. C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformaitcs* 34, 33–40. doi: 10.1093/bioinformatics/btx579

Liu, B., Liu, F., Fang, L., Wang, X., and Chou, K. C. (2015b). repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics* 31, 1307–1309. doi: 10.1093/bioinformatics/btu820

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K. C. (2015c). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Liu, B., Wang, S., Long, R., and Chou, K. C. (2017a). iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* 33, 35–41. doi: 10.1093/bioinformatics/btw539

Liu, B., Wu, H.,Zhang, D., Wang, X., and Chou, K. C. (2017b). Pse-analysis: a python package for DNA, RNA and protein peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget* 8, 13338–13343. doi: 10.18632/oncotarget.14524

Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., et al. (2014). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 30, 472–479. doi: 10.1093/bioinformatics/btt709

Liu, L. M., Xu, Y., and Chou, K. C. (2017). iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. *Med. Chem.* 13, 552–559. doi: 10.2174/1573406413666170515120507

Liu, X., Yu, Y., Liu, J., Elliott, C. F., Qian, C., and Liu, J. (2018). A novel data structure to support ultra-fast taxonomic classification of metagenomic sequences with k-mer signatures. *Bioinformatics* 34, 171–178. doi: 10.1093/bioinformatics/btx432

Luo, L., Li, D., Zhang, W., Tu, S., Zhu, X., and Tian, G. (2016). Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features. *PloS ONE* 11:e0153268. doi: 10.1371/journal.pone.0153268

Matias Rodrigues, J. F., Schmidt, T. S. B., Tackmann, J., and von Mering, C. (2017). MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* 33, 3808–3810. doi: 10.1093/bioinformatics/btx517

Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *J. Electr. Imaging* 16:049901. doi: 10.18637/jss.v017.b05

Ni, P. X., Dai, W. K., Liu, Y. F., Yang, Z. Y., Zhou, T., Liang, S. Q., et al. (2017). A novel method for better bacterialgenome assembly from illumina data. *Curr. Bioinform.* 12, 498–508. doi: 10.2174/1574893610666150624171516

Orenstein, Y., Pellow, D., Marçais, G., Shamir, R., and Kingsford, C. (2017). Designing small universal k-mer hitting sets for improved analysis of high-throughput sequencing. *PLoS Comput. Biol.* 13:e1005777. doi: 10.1371/journal.pcbi.1005777

Pan, Y., Liu, D., and Deng, L. (2017a). Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. *PloS ONE* 12:e0179314. doi: 10.1371/journal.pone.0179314

Pan, Y., Wang, Z., Zhan, W., and Deng, L. (2017b). Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 34, 1473–1480. doi: 10.1093/bioinformatics/btx822

Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526, 519–524. doi: 10.1038/nature14666

Qiao, Y., Xiong, Y., Gao, H., Zhu, X., and Chen, P. (2018). Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. *BMC Bioinform.* 19:14. doi: 10.1186/s12859-018-2009-5

Qiu, W. R., Sun, B. Q., Xiao, X., Xu, Z. C., and Chou, K. C. (2016). iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics* 32, 3116–3123. doi: 10.1093/bioinformatics/btw380

Qiu, W. R., Xiao, X., Lin, W. Z., and Chou, K. C. (2015). iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J. Biomol. Struct. Dyn.* 33, 1731–1742. doi: 10.1080/07391102.2014.968875

Rangavittal, S., Harris, R. S., Cechova, M., Tomaszkiewicz, M., Chikhi, R., Makova, K. D., et al. (2018). RecoverY: k-mer-based read classification for Y-chromosome-specific sequencing and assembly. *Bioinformatics* 34, 1125–1131. doi: 10.1093/bioinformatics/btx771

Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J. M., Kim, J., et al. (2017). Recurrent and functional regulatory mutations in breast cancer. *Nature* 547, 55–60. doi: 10.1038/nature22992

Senawi, A., Wei, H. L., and Billings, S. A. (2017). A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. *Pattern Recogn.* 67, 47–61. doi: 10.1016/j.patcog.2017.01.026

Shipman, S. L., Nivala, J., Macklis, J. D., and Church, G. M. (2017). CRISPR–cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* 547, 345–349. doi: 10.1038/nature23017

Sleight, S. C., Bartley, B. A., Lieviant, J. A., and Sauro, H. M. (2010). In-fusion biobrick assembly and re-engineering. *Nucleic Acids Res.* 38, 2624–2636. doi: 10.1093/nar/gkq179

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics*. doi: 10.1093/bioinformatics/bty508. [Epub ahead of print].

Tang, H., Chen, W., and Lin, H. (2016). Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Mol. BioSyst.* 12, 1269–1275. doi: 10.1039/c5mb00883b

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Thomas, C. A. Jr. (1971). The genetic organization of chromosomes. *Annu. Rev. Genet.* 5, 237–256.

Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., et al. (2013). Frequency of TERT promoter mutations in human cancers. *Nat. Commun* 4:2185. doi: 10.1038/ncomms3185

Vogel, F. (1964). A preliminary estimate of the number of human genes. *Nature* 201:847.

Warrens, A. N., Jones, M. D., and Lechler, R. I. (1997). Splicing by overlap extension by PCR using asymmetric amplification: an improved technique for the generation of hybrid proteins of immunological interest. *Gene* 186, 29–35.

Wei, H. L., and Billings, S. A. (2007). Feature subset selection and ranking for data dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 162–6. doi: 10.1109/TPAMI.2007.11

Wei, L., Liao, M., Gao, X., and Zou, Q. (2015). Enhanced protein fold prediction method through a novel feature extraction technique. *IEEE Trans. Nanobiosci.* 14, 649–659. doi: 10.1109/TNB.2015.2450233

Wu, Y., Li, B. Z., Zhao, M., Mitchell, L. A., Xie, Z. X., Lin, Q. H., et al. (2017). Bug mapping and fitness testing of chemically synthesized chromosome X. *Science* 355:eaaf4706. doi: 10.1126/science.aaf4706

Xiao, Y., Zhang, J., and Deng, L. (2017). Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. *Sci. Rep.* 7:3664. doi: 10.1038/s41598-017-03986-1

Xie, Z. X., Li, B. Z., Mitchell, L. A., Wu, Y., Qi, X., Jin, Z., et al. (2017). "Perfect" designer chromosome V and behavior of a ring derivative. *Science* 355:eaaf4704. doi: 10.1126/science.aaf4704

Xiong, Y., Liu, J., and Wei, D. Q. (2011). An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins* 79, 509–517. doi: 10.1002/prot.22898

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., et al. (2017). PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.* 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Yang, H., Qiu, W. R., Liu, G. Q., Guo, F. B., Chen, W., Chou, K. C., et al. (2018). iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.* 14, 883–891. doi: 10.7150/ijbs.24616

Yang, H., Tang, H., Chen, X. X., Zhang, C. J., Zhu, P. P., Ding, H., et al. (2016). Identification of secretory proteins in *Mycobacterium* tuberculosis using pseudo amino acid composition. *Biomed. Res. Int.* 2016:5413903. doi: 10.1155/2016/5413903

Yao, Y. H., Li, X. H., Geng, L. L., Nan, X. Y., Qi, Z. H., and Liao, B. (2018). Recent progress in long noncoding RNAs prediction. *Curr. Bioinformatics* 13, 344–351. doi: 10.2174/1574893612666170905153933

You, Z. H., Huang, Z. A., Zhu, Z., Yan, G. Y., Li, Z. W., and Wen, Z. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005455. doi: 10.1371/journal.pcbi.1005455

Zhang, N., Yu, S., Guo, Y., Wang, L., Wang, P., and Feng, Y. (2018). Discriminating Ramos and Jurkat Cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinform.* 13, 50–6., doi: 10.2174/1574893611666160608102537

Zhang, W., Bojorquez-Gomez, A., Velez, D. O., Xu, G., Sanchez, K. S., Shen, J. P., et al. (2018). A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* 50, 613–620. doi: 10.1038/s41588-018-0091-2

Zhang, W., Liu, J., Zhao, M., and Li, Q. (2012a). Predicting linear B-cell epitopes by using sequence-derived structural and physicochemical features. *Int. J. Data Min. Bioinform.* 6, 557–569. doi: 10.1504/IJDMB.2012.049298

Zhang, W., Niu, Y., Xiong, Y., Zhao, M., Yu, R., and Liu, J. (2012b). Computational prediction of conformational B-cell epitopes from antigen primary structures by ensemble learning. *PloS ONE* 7:e43575. doi: 10.1371/journal.pone.0043575

Zhang, W., Niu, Y., Zou, H., Luo, L., Liu, Q., and Wu, W. (2015). Accurate prediction of immunogenic T-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning. *PloS ONE* 10:e0128194. doi: 10.1371/journal.pone.0128194

Zhang, W., Zhao, G., Luo, Z., Lin, Y., Wang, L., Guo, Y., et al. (2017b). Engineering the ribosomal DNA in a megabase synthetic chromosome. *Science* 355:eaaf3981. doi: 10.1126/science.aaf3981

Zhang, W., Zhu, X., Fu, Y., Tsuji, J., and Weng, Z. (2017a). Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. *BMC Bioinform.* 18(Suppl. 13):464. doi: 10.1186/s12859-017-1875-6

Zhou, L. Q., Li, R., and Hu, L. (2016). Enhanced prediction of small non-coding RNA in bacterial genomes based on improved inter-nucleotide distances of genomes. *Curr. Bioinform.* 11, 169–72. doi: 10.2174/1574893611666160223201114

Zou, Q., Liu, W., Merler, M., and Ji, R. (2016). Advanced learning for large-scale heterogeneous computing. *Neurocomputing* 217, 1–2. doi: 10.1016/j.neucom.2016.06.009

Conflict of Interest Statement: 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.

Copyright © 2018 He, Ju, Zeng, Liu and Zou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

TECHNOLOGY REPORT
published: 08 October 2018
doi: 10.3389/fmicb.2018.02322

PanGFR-HM: A Dynamic Web Resource for Pan-Genomic and Functional Profiling of Human Microbiome With Comparative Features

*Narendrakumar M. Chaudhari[1†], Anupam Gautam[1,2†], Vinod Kumar Gupta[1‡], Gagneet Kaur[1,2], Chitra Dutta[1] and Sandip Paul[1]**

[1] *Structural Biology and Bioinformatics Division, CSIR-Indian Institute of Chemical Biology, Kolkata, India,* [2] *Department of Pharmacoinformatics, National Institute of Pharmaceutical Education and Research, Kolkata, India*

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Yazhou Sun,
Shenzhen University, China
Wen Zhang,
Chinese Center for Disease Control and Prevention, China

***Correspondence:**
Sandip Paul
sandippaul@iicb.res.in;
websandip@gmail.com

[†] *These authors have contributed equally to this work*

[‡] ***Present address:***
Vinod Kumar Gupta,
Mayo Clinic, Rochester, MN, United States

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

Received: *10 May 2018*
Accepted: *11 September 2018*
Published: *08 October 2018*

Citation:
Chaudhari NM, Gautam A, Gupta VK, Kaur G, Dutta C and Paul S (2018) PanGFR-HM: A Dynamic Web Resource for Pan-Genomic and Functional Profiling of Human Microbiome With Comparative Features. Front. Microbiol. 9:2322. doi: 10.3389/fmicb.2018.02322

The conglomerate of microorganisms inhabiting various body-sites of human, known as the human microbiome, is one of the key determinants of human health and disease. Comprehensive pan-genomic and functional analysis approach for human microbiome components can enrich our understanding about impact of microbiome on human health. By utilizing this approach we developed PanGFR-HM (http://www.bioinfo.iicb.res.in/pangfr-hm/) – a novel dynamic web-resource that integrates genomic and functional characteristics of 1293 complete microbial genomes available from Human Microbiome Project. The resource allows users to explore genomic/functional diversity and genome-based phylogenetic relationships between human associated microbial genomes, not provided by any other resource. The key features implemented here include pan-genome and functional analysis of organisms based on taxonomy or body-site, and comparative analysis between groups of organisms. The first feature can also identify probable gene-loss events and significantly over/under represented KEGG/COG categories within pan-genome. The unique second feature can perform comparative genomic, functional and pathways analysis between 4 groups of microbes. The dynamic nature of this resource enables users to define parameters for orthologous clustering and to select any set of organisms for analysis. As an application for comparative feature of PanGFR-HM, we performed a comparative analysis with 67 *Lactobacillus* genomes isolated from human gut, oral cavity and urogenital tract, and therefore characterized the body-site specific genes, enzymes and pathways. Altogether, PanGFR-HM, being unique in its content and functionality, is expected to provide a platform for microbiome-based comparative functional and evolutionary genomics.

Keywords: comparative genomics, database, functional profile, human microbiome, pan genome, web resource

INTRODUCTION

The variety of microorganisms inhabiting different body-sites of human – is one of the key determinants of human health and disease. Recent emergence of metagenomic approaches, empowered by the technical and conceptual advancements in low-cost, high-throughput sequencing methodologies have enabled the scientific community to understand the genetic/functional diversity of the "healthy microbiome" components, a crucial step for identifying the microbial species that are implicated in disease (Reid et al., 2011; Gupta et al., 2017). The vast resource of microbial reference genomes from different body-sites of healthy humans, available at Data Analysis and Coordination Center of Human Microbiome Project (HMP-DACC)[1], provides the scientific community an opportunity to comprehend the genomic landscape and thus functional potential of any particular group of organisms in various body habitats (NIH HMP Working Group et al., 2009; Human Microbiome Project Consortium, 2012).

One of the major bioinformatic frameworks that have been proven to be useful and informative in comparative analysis of multiple microbial genomes is the 'pan-genome' approach developed by Tettelin et al. (2005). Pan-genome of a given species/taxon represents the complete set of non-redundant genes from its representative genomes and is comprised of three parts: core genes (representatives from all genomes), accessory genes (representatives from two or more genomes, not all) and genome specific genes. The pan-genomic profiling and subsequent systemic functional annotation at various taxonomic levels, varying from within-species community to cross-species communities at intra-/inter-habitat level, offer evolutionary insights and potential functional importance of any group of microorganisms. Moreover, various reports reveal that the comparative pan-genome analysis has tremendous potential for offering new perspective on the species diversity and adaptive strategies of human microbiome in body-site specific manner (Rasko et al., 2008; Conlan et al., 2012; Gupta et al., 2015; Bakshi et al., 2016; Duranti et al., 2016). Therefore, a comprehensive resource of human microbiome providing in-depth pan-genomic analysis of strains at various taxonomic levels, with subsequent estimation of the functional repertoire from same or different body-sites along with comparative analysis approach will be of great interest.

The existing database tools for pan-genome analysis of microbes like, MetaRef (Huang et al., 2014), MicroScope platform (Vallenet et al., 2017), and EDGAR 2.0 (Blom et al., 2016) provide basic pan-genomic information about the microbes in general but lack features like user-defined selection of strains or isolation body-site from human, in-depth strain wise pan-genomic details of shared genes, and strain specific presence/absence of genes along with their functional profiling. Also, there is no such resource which allows users to investigate/compare pan-genomes of multiple user defined groups within human microbiome strains. To this end, we developed PanGFR-HM – Pan-Genomic and Functional Repertoire of Human Microbiome components – an online dynamic resource that systematically integrates the functional and compositional characteristics of complete gene repertoire of 1293 reference bacterial and archaeal genomes from HMP-DACC. It offers options for pan-genomic analysis, potential functional analysis using Clusters of Orthologous Genes (COG) and Kyoto Encyclopedia of Genes and Genomes (KEGG), and comparative analyses for any possible combinations of genomes. The features for pan-genome analysis provide information about core, accessory and unique gene families among a user defined set of genomes, which can belong to a specific taxonomical clade or body-site. PanGFR-HM allows the users to explore the genomic and functional diversity, potential lateral gene transfer events and phylogenetic relationships between human associated microbial genomes, which are not provided by any existing public domain computational resources. Exceptionally, within a user defined set of genomes, this resource provides information about probable gene loss events, i.e., the genes exclusively absent from a specific genome but present in all other genomes. Also, significant over/under representation of KEGG/COG functional categories in different gene families (core, accessory, unique) are provided for that dataset. Most importantly, this resource enables users to perform comparative analysis between different groups of microbes (based on taxonomy and/or body-site) for common as well as group specific functional and gene-family architectures. All the results can be accessed freely through an online web-interface, interactively and can be downloaded for further analysis. We envision that, PanGFR-HM, being unique in its content and functionality, will greatly facilitate the progress of microbiome-based evolutionary research, clinical application of microbial genomics and create footprints for future studies on the composition-activity relationship of the human microbiome components.

MATERIALS AND METHODS

Overview of PanGFR-HM

PanGFR-HM serves as an ample and appropriate resource for exploring the genomic and functional repertoire and diversity, phylogenetic relationships among human associated microbial genomes by providing numerous attributes not available in any existing computational resources. All 1293 strains belong to 8 major bacterial/archaeal phyla, i.e., *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Fusobacteria*, *Proteobacteria*, *Spirochaetes*, *Synergistetes,* and *Euryarchaeota*. At genus level, these genomes represent 187 different defined genera (see **Supplementary Table S1**). These microbes, as part of human microbiome, comprise mostly of bacteria derived from distinct body-sites of human (Detailed list of microbial species is provided in **Supplementary Table S1**). Gene families (gene clusters) generated from all annotated proteins from complete genomes of these microbes were integrated into a database, where pan-genomic details of any subset of these microbial strains belonging to specific taxonomical clade or body-site, can be dynamically retrieved.

[1]https://www.hmpdacc.org/HMRGD/

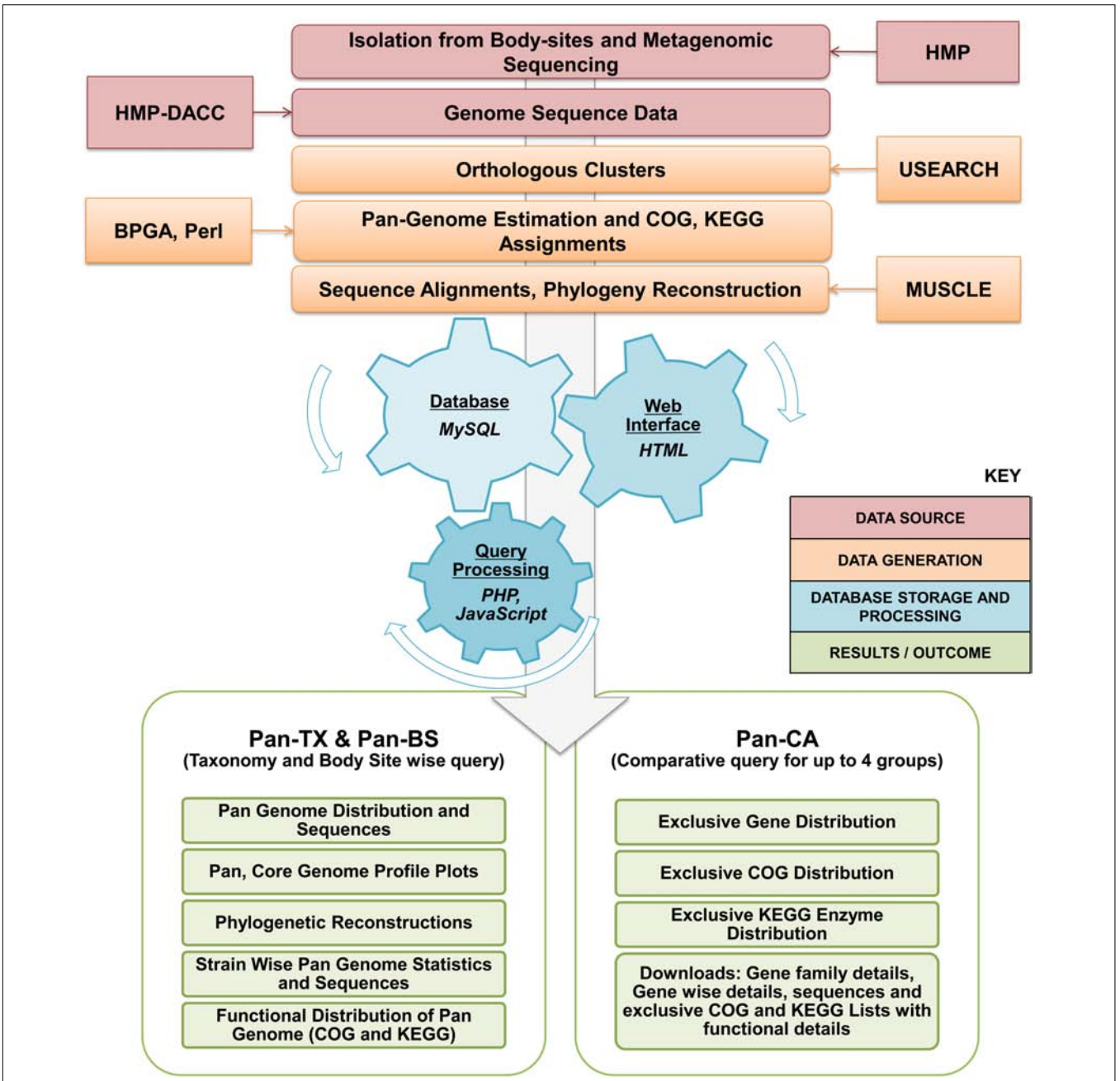


FIGURE 1 | Workflow for PanGFR-HM. The workflow involves preprocessing of the whole genome records from GenBank. The protein sequences from all the available strains are then clustered using USEARCH at various sequence identity cut off levels. The gene presence/absence is determined and stored as a binary matrix using BPGA Pipeline. The tabular data generated after sequence processing and functional assignments are then uploaded into separate MySQL databases for respective sequence identity cut offs. Majority of the backend programs are written in PHP. Finally, various pan genomic and comparative analysis results generated from the set of strains and clustering parameters selected by the users are then provided via the web browser.

PanGFR-HM provides the pan-genomic profile for genomes of the interest based on user defined sequence identity criteria for protein sequences (ranging from 40 to 90%) for detection of orthologous clusters. The pan-genome profile comprises of comprehensive information about core gene families, accessory gene families, gene families with genome wise exclusive presence and absence, and prediction of nature of pan-genome (open/close) with statistics. PanGFR-HM integrates additional features for reconstructing the phylogenetic relationships among selected genomes based on concatenated core genes (users can select 10, 20, 30, 50, 70, or 100 random core genes for this purpose, 20 by default) as well as gene presence/absence profile (pan-genome tree). PanGFR-HM can provide the functional composition (based on COG and KEGG annotations) of

core, accessory and unique gene families with over/under representation statistics for genomes of the interest. It is also capable of delivering information about the genes exclusively absent from a specific genome but present in all other genomes within a group, indicating probable gene loss events. Apart from these, another important feature is Pan-CA, which enables users to perform the comparative analyses of pan-genomes and function/pathway annotations of core, accessory and unique genes for up to four user defined groups of pan-genomes.

The web interface for PanGFR-HM has been developed to offer a user-friendly way to access the taxonomic and body-site specific interactive view to explore the divergence in gene repertoire and functional composition among human microbiota. The resource utilizes latest plotting, data storage and computing libraries from various free community resources. All information, including pan-genome profiles, phylogenetic trees (based on both concatenated core genes and gene presence/absence profile), COG and KEGG annotation distribution (for core, accessory and unique gene families), and protein sequences (core, accessory, unique and genes exclusively absent from a particular strain) incorporated in PanGFR-HM are available for download in publication level graphical, tree (newick), table (xls) and text (fasta format of sequences) formats wherever applicable. The protein sequences can be downloaded as representative sets for core/accessory/unique gene families as well as for all the members of each gene family. These sequence files can easily be used further for evolutionary analyses, domain/motif search, study of physicochemical properties etc. PanGFR-HM not only provides novel aspects such as body-site specificity and comparative analysis, but also allows users to choose the genomes of their interest as well as sequence identity criteria for orthology detection. The different levels of sequence identity for orthology prediction allow users to precisely target various evolutionary distances within human microbiota (Pearson, 2013). These features provide PanGFR-HM a 'dynamic' status instead of 'static' database unlike MetaRef, MicroScope platform and EDGAR 2.0 where, no such user defined options are available. PanGFR-HM is the only dynamic database especially dedicated to human microbiome and integrated huge information with unique functionality compared to its analogs.

Database Design, Organization and Structure

The PanGFR-HM logistics has been shown schematically in **Figure 1**. The detailed schema for the database and its connections to the web resource is available in **Supplementary Figure S1**. The resource integrates bacterial and archaeal reference genome data derived from human microbiome and delivers the outcome in the form of pan-genome profile. An easy to use web interface allows users to retrieve the pan-genomic profile and information of functional distribution for any set of available genomes.

For a user defined set of genomes the extrapolation of pan and core genome curves can be performed by empirical power law equations and exponential decay equations respectively as implemented by Bacterial Pan-Genome Analysis Pipeline (BPGA) (Chaudhari et al., 2016). Slope of the power curve (the *B* value), helps users to decide the open/closed nature of pan-genome, i.e., whether the pan-genome size increases considerably after inclusion of additional microbial genome or the saturation is achieved. Phylogenetic analysis can be retrieved from core orthologous clusters and binary presence/absence matrix (pan matrix) using MUSCLE (Edgar, 2004). It first aligns the concatenated protein sequences of core proteins and then builds Neighbor Joining tree upon the alignment. Users can select the number of random core proteins (10, 20, 30, 50, 70, and 100 – default 20) in order to reconstruct the phylogenetic tree. If less number of core proteins than the user-defined core proteins are present, all of them will be considered for phylogenetic tree reconstruction. The overall topology of this random core-genome tree remains unaltered as compared to the tree formed using all core protein sequences when present in large number (Chaudhari et al., 2016). Core, accessory and unique protein families are then assigned for given set of genomes along with their sequences and function/pathway annotations. The functions are annotated using NCBI COG database, 2014 update (Galperin et al., 2015) and KEGG enzymes are annotated using KAAS server (Moriya et al., 2007).

The home page of PanGFR-HM serves as the gateway to the interlinked genomic and functional features. The interface is capable of utilizing the database features dynamically as instructed through interactive web input forms at the respective web modules. The web resource is compatible with the latest versions of Edge (version 41+), Google Chrome (version 66.0+), Safari (version 11.1+), and Mozilla Firefox (version 59.02+).

Data Generation

The high quality complete genome sequences for 1293 bacteria and archaea were downloaded from HMRGD (HMP Reference Genome sequence Data)[1]. The protein sequences and annotations were extracted from the GenBank records for the same. Protein sequences were clustered separately into orthologous gene families at different sequence identity cut-off values of 40, 50, 60, 70, 80, and 90% using USEARCH (Edgar, 2010). The orthologous clusters were then processed using BPGA (Chaudhari et al., 2016). Using the features of BPGA pipeline, paralogs were discarded for the ease of analysis and binary gene presence/absence matrix was generated. Each orthologous cluster was then mapped with latest NCBI COG database (last updated 2014)[2] using best blast-hits for annotation of functions and then the assignments of pathways were done by KAAS v2.1 (KEGG Automatic Annotation Server)[3] using BBH (bi-directional best hit) method using representative protein sequences (Moriya et al., 2007; Galperin et al., 2015).

Database Creation

All the clustering data along with sequence and function data were integrated into MySQL community database engine (v5.7)

[2]ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data/

[3]http://www.genome.jp/tools/kaas/

in organized manner for each identity cut-off level so that the orthology data can be retrieved based on a query provided by the users.

Data Processing and Delivery

Web pages were designed in HTML5. User forms and all other calculations including SQL database queries were processed in PHP (v7.0.9) and JavaScript. Most of the plots generated during these analyses used Plotly (v1.29.1), the open source JavaScript graphing library[4]. Sequence alignments and phylogeny trees were generated using MUSCLE (v3.8.31) (Edgar, 2004). Users can also import the phylogenetic trees to iTOL (Interactive Tree Of Life) web server[5] for better visualizations, formatting and high resolution graphics (Letunic and Bork, 2016). Phylocanvas Library is used for interactive tree visualizations[6].

Characterization of Pan-Genome

Pan-genome characterization of group of genomes is a dynamic process and depends upon the criteria for construction of orthologous gene families or clusters generated from clustering tools. We utilized the USEARCH clustering tool (Linux v9.2.64) for all proteins from 1293 currently accessible reference genomes derived from human microbiome at HMRGD[1]. Using PanGFR-HM web form, users can select any number of genomes (maximum 200 genomes recommended) either body-site wise or taxonomy-wise for an analysis, and consider any of the amino acid identity cut-offs (ranging from 40 to 90% with steps of 10) for estimating the orthologous clusters. On the basis of selected identity cut-off value, the respective protein families are then extracted from database along with sequence and functional details to build the pan-genomic and functional profile.

Functional Over/Under Representation Analysis

For a group of genomes, the differentially represented functional sub categories of each major category of COG and KEGG classification for pan-genome component (core, accessory and unique) proteins are determined based on the respective major category as reference. The statistical analysis for the significance testing is performed using Chi-Square Test with 1 degree of freedom. The following formula is used for calculation of Chi-Square value for a particular sub category within a major category of a specific pan-genome component,

$$x^2 = \frac{n \cdot (a \cdot b - b \cdot c)^2}{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}$$

Where, $n = a + b + c + d$; a is the count of COG/KEGG assignments of that particular functional sub category and b is the count of the rest of that sub categories of that specific pan-genome component, c and d are the respective counts of COG/KEGG assignments of same functional sub category and rest of the sub categories of remaining two pan-genome components. The functional sub categories which pass the significance test are marked accordingly for over or under representation.

Methodology for Comparative Analysis

In Pan-CA module the comparative gene analysis is performed in two steps. First, the orthologous gene clusters from all member genomes of each group selected by users are identified and next every possible shared and exclusive gene clusters between the groups are calculated. For example, if users select strains for three groups (A, B, and C) then total seven possible sets will be there: one core set (ABC), three accessory sets (AB, AC, and BC) and three unique sets (A, B, and C). Further the COG/KEGG classification of shared and exclusive gene clusters is presented in both graphical and tabular format. For comparative function analysis and comparative pathway analysis in Pan-CA, only the annotated COG protein identifiers and KEGG enzyme identifiers of all the selected genomes are extracted and pooled instead of gene clusters, followed by group-wise comparison for shared and exclusive COG/KEGG identifiers. All the results are then presented by plotting Venn diagrams (downloadable SVG or PNG images) and providing tabular output with browsing options and downloadable links.

RESULTS

Data Overview and Statistics

The pan-genome statistics of selective genera of human microbiome present in PanGFR-HM are summarized in **Figure 2**. The genera containing at least 5 complete genomes are selected for this analysis. Along with core, accessory and unique gene counts, the figure also depicts the *B* statistic of each pan-genome at both 50 and 80% amino acid sequence identity cut-offs. *B* statistic value gives an idea about the open or closed nature of pan-genome. The *B* value toward '1' indicates the open pan-genome where pan-genome size constantly rises after stepwise addition of new genomes. Whereas, the *B* value toward '0' indicates closed pan-genome where pan-genome size does not change after inclusion of additional genomes.

As shown in **Figure 2**, pan-genomes of the genera *Aeromonas, Finegoldia, Mobiluncus, Myroides*, *Peptoclostridium,* and *Rothia* seem to add fewer new genes with addition of new genomes with *B* value < 0.4 (Chaudhari et al., 2016). These estimates may be misleading as they are based on predictions from only few available members of a genus (only 5–8 genomes). Whereas, the pan-genomes of genera *Escherichia*, *Klebsiella*, *Propionibacterium* and *Staphylococcus* are found to be not growing rapidly with lower *B* values of 0.3/0.33, 0.35/0.38, 0.26/0.31, and 0.39/0.41 based on 35, 14, 80, and 56 genomes at 50/80% sequence identity cut-offs, respectively.

Query Options

PanGFR-HM offers various features for flexible query and comprehensive pan-genomics as well as comparative

[4]https://plot.ly/javascript/

[5]https://itol.embl.de/

[6]http://phylocanvas.org/

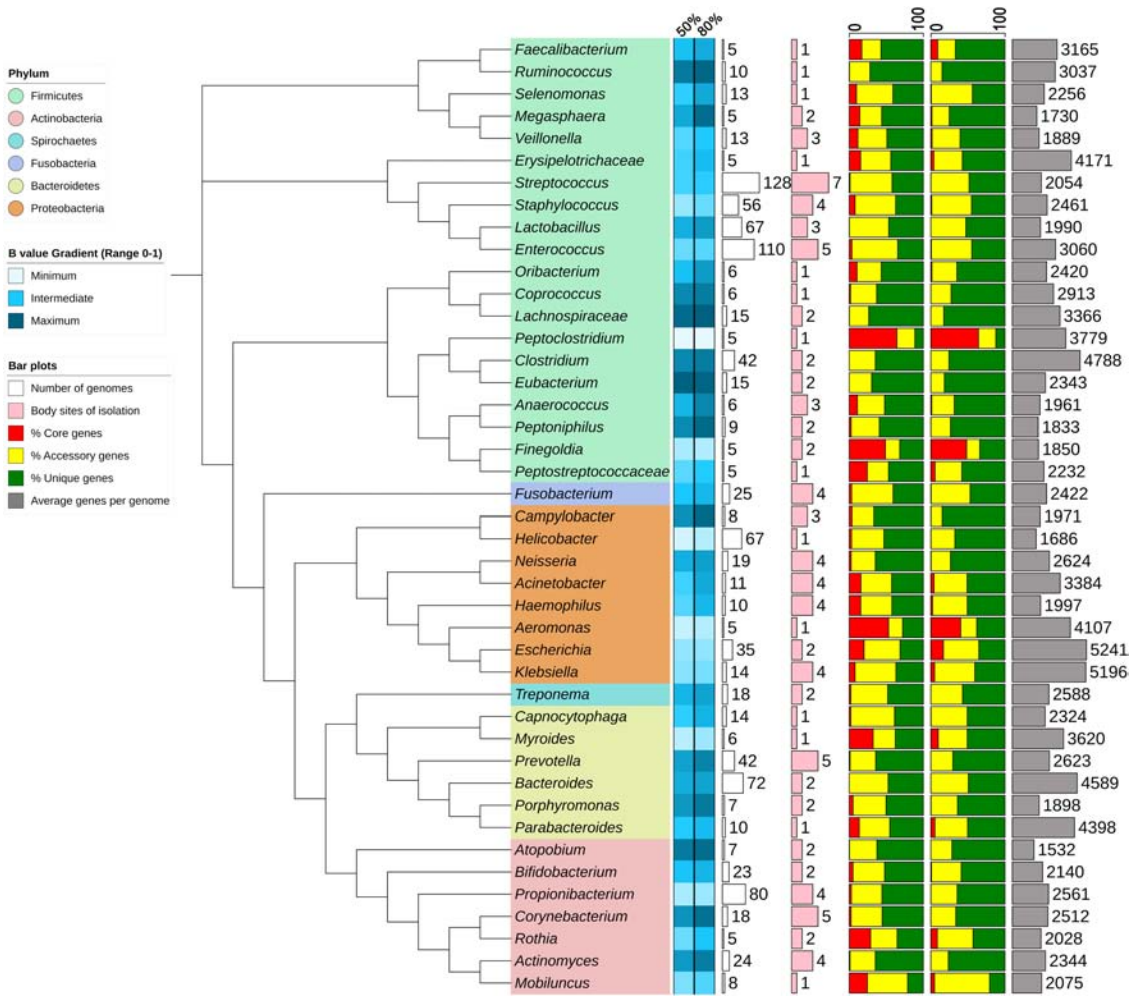


FIGURE 2 | Summary of pan-genome statistics for selected genera. The Neighbor Joining phylogenetic tree is constructed using 16S rRNA genes from representatives of each taxon. The *B* values are depicted as heatmap and the percentages of core, accessory and unique genes as stacked bar plots for each taxon using 50 and 80% sequence identity cut-offs. The plot and heatmap are constructed by iTOL web server using statistics calculated from PanGFR-HM.

analysis of human microbiome strains. The resource can be navigated through any of the three options: (I) Taxonomy-wise Pan-Genome and Functional Analysis, (II) Body-site wise Pan-Genome and Functional Analysis, and (III) Comparative Pan-Genome and Functional Analysis for flexible and rational selection of strains based on various criteria. All of them deliver in-depth analysis of the genomic and functional repertoire of selected strains. Apart from these we have also integrated the BLAST[7] program within this resource. Therefore users can perform BLAST search for their query sequences against any pan-genomic profile of group of genomes.

The performance of this resource mainly depends on the size of the selected dataset by users and collective server load. The resource took around 15 min for pan-genomic analysis of top 10 genera (based on number of strains present) having total 699 strains run in parallel.

[7]https://blast.ncbi.nlm.nih.gov/Blast.cgi

Taxonomy-Wise Pan-Genome and Functional Analysis (Pan-TX)

This module enables pan-genomic analysis of any set of the available strains from HMP based on their taxonomy. The users can select all the human microbiome strains from a desired species, genus or any other taxonomic level irrespective of the isolation site within human body. It provides phylogenetic tree reconstruction of the selected strains based on the approaches like pan-genome (gene presence/absence) and core-genome (concatenated and aligned amino acid sequences of core genes) along with the comprehensive pan-genomic and potential functional repertoire of selected taxon.

Body-Site Wise Pan-Genome and Functional Analysis (Pan-BS)

This module enables users to select microbiome strains for analysis on the basis of their major site of isolation within human body as defined by HMP. Users can also select only

the strains isolated from a particular body-site to extract information about gene, function and pathway repertoire among the selected strains along with the routine pan-genomic analysis results.

Comparative Pan-Genomic and Functional Analysis (Pan-CA)

The Pan-CA module is another flexible and novel feature of PanGFR-HM. This module enables users to make a flexible query for analysis of up to 4 distinct groups of strains and derive the comparative picture of genes, functions and pathways among selected groups (pan-genomes). The groups can be formed on the basis of taxonomy (like Pan-TX), isolation site of microbes (like Pan-BS) or any other suitable criteria decided by the users.

Output Options

Pan-genome analysis performed on strains of interest, selected via Pan-TX or Pan-BS, delivers comprehensive pan-genome and functional analyses results. The results include: details of selected strains (dataset), overall pan-genome statistics (proportion of core, accessory, unique genes) for given set of genomes, core and pan-genome profile plots, phylogenetic reconstruction based on core genes and pan-genome, genes specifically absent from individual strain, distribution of proteins in different COG and KEGG functional categories and their over/under representation for each pan-genomic component, and strain wise pan-genome statistics along with data or sequence download links for all plots, phylogenetic trees and protein sequences etc.

The comparative analyses performed in Pan-CA module on groups of microbiome strains of interest provide results for orthologous proteins, COG identifiers and KEGG enzyme identifiers for all possible sets (shared and unique) between up to four groups. Distribution of proteins or identifiers in every possible set is explained with Venn diagrams, and data for each of these sets is provided as spreadsheets. For comparative analysis of orthologous proteins downloadable FASTA sequences for further analyses and COG/KEGG classification details with plots are also given.

When BLAST search is performed with protein sequences uploaded by users, it generates mainly two kinds of outputs. One of them includes pan-genomic distribution plot of gene clusters from selected strains for building the database. The other depicts the BLAST output spreadsheet showing how many proteins among the queried proteins have pan-genomic orthologs along with pan-genomic status (core/accessory/unique), KEGG identifiers, COG identifiers, sequence alignment details etc. For each orthologous proteins clickable links are given to corresponding alignments, COG/KEGG and gene identifiers details. Also a distribution plot is available summarizing pan-genomic distribution of orthologous proteins.

Additional Novel Features

Dynamic Estimation of Pan-Genome

Pan-genome characterization of a group of genomes is a dynamic process, which greatly depends upon the criteria for construction of orthologous gene families or clusters generated from the sequence clustering tools. The pan-genome estimation may highly fluctuate for different sequence identity cut-off criteria depending upon the rate of divergence, although overall pan-genome characteristic does not vary much for closely related genomes (Paul et al., 2016). Using PanGFR-HM web form, users may select at least 5 to all genomes (maximum 200 recommended, for more than that the resource will take longer time) at a time, and proceed for analysis based on various sequence identity cut-offs ranging from 40 to 90% for constructing orthologous protein clusters. This feature brands PanGFR-HM as a dynamic server, not just a static database with pre-calculated clusters with fixed parameters.

TABLE 1 | Summary dataset of *Lactobacillus* strains used for comparative analysis.

Body site	No. of strains	Total proteins	Average proteins per genome
Gastrointestinal tract	23	53687	2334
Oral cavity	4	10534	2634
Urogenital tract	40	69081	1727

Exclusive Absence of Genes: A Clue to Gene Loss Events

It is well known that bacterial genomes acquire new genes from surrounding gene pools to get an adaptive advantage to the environmental or cellular changes (Dutta and Pan, 2002; Popa et al., 2011; Arber, 2014; Li et al., 2014). Most of these genes fall under unique genes category in any pan-genome analysis due to lack of orthologs in related organisms. Apart from these unique genes, another very important evolutionary process is gene loss, which may be another adaptive strategy for genome evolution (Hottes et al., 2013; Bolotin and Hershberg, 2015). The gene loss events are often hard to track down at sequence level. A novel feature is integrated in PanGFR-HM for investigating the genes exclusively absent (not matching under given sequence identity cut off) from a genome but present in all other genomes of the users selected dataset. By exclusive gene absence analysis in PanGFR-HM, one can estimate such probable events, *in silico*. These exclusively absent genes might also be important for adaptation of the microbes at a specific niche. PanGFR-HM specifically extracts those gene families and provides their sequences for download and function annotations.

Functional Over/Under Representation Analysis

The assignments of COG and KEGG functional classification are done for core, accessory and unique gene sets. The significantly over and underrepresented functional categories within a major category among the above sets are reported. The feature aids in understanding the gene divergences which led to the functional evolution of pan-genome.

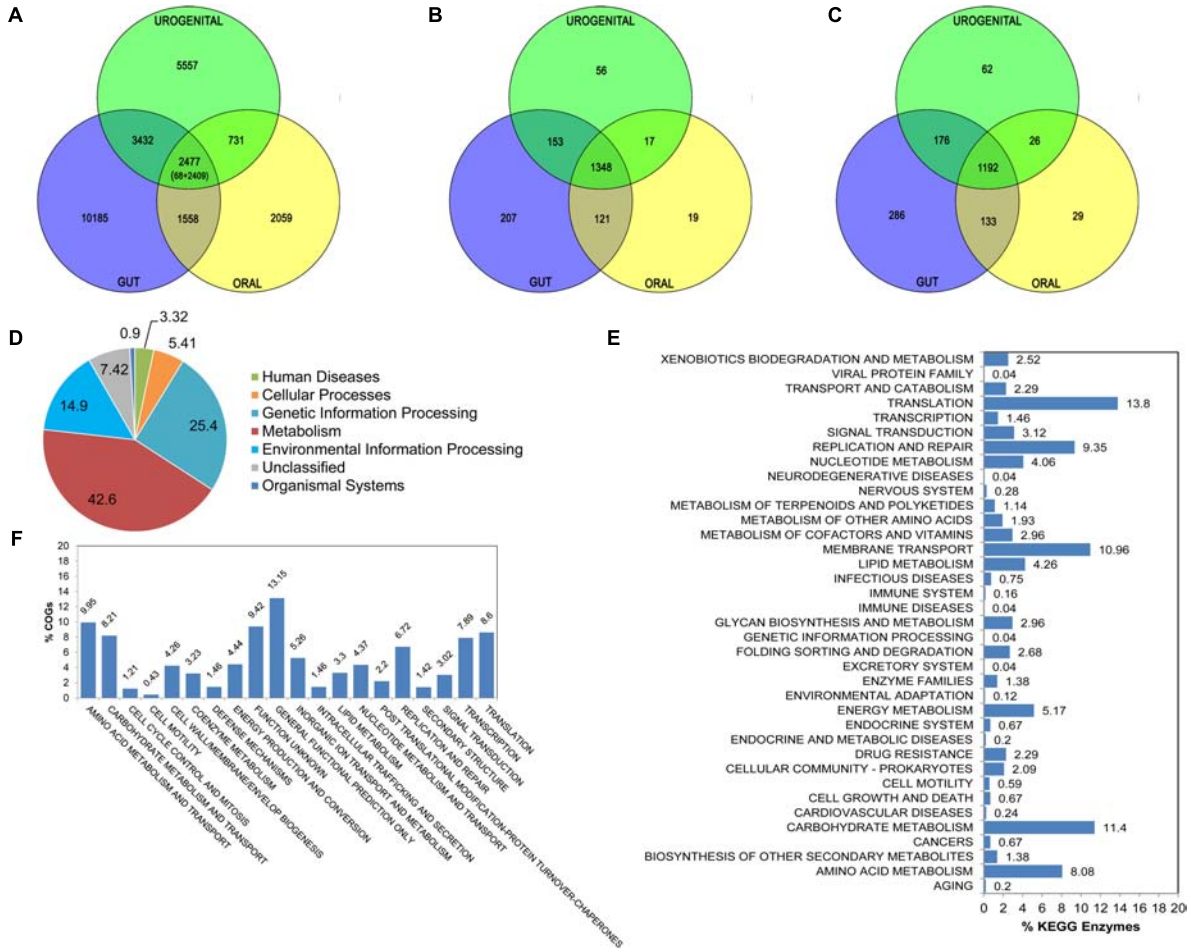


FIGURE 3 | Comparative analysis for *Lactobacillus* strains from gut, oral cavity and urogenital tracts. Venn diagrams depict the number of gene families **(A)**, COG proteins **(B)** and KEGG enzymes **(C)** at three body-sites. Distribution of major KEGG functional categories for 1192 core enzymes is shown as pie chart **(D)**. Bar plots are created in order to illustrate the distribution of detailed core functional KEGG enzymes **(E)** and COG proteins **(F)**.

BLAST Search Against Pan-Genomic Profile

This feature allows users to paste/upload their own protein sequences in FASTA format and perform the BLAST search against user defined pan-genomic profile from PanGFR-HM. Users have the option to select strains of interest (either based on taxonomy or isolation site) in order to create a representative set of pan-genomic profile, which will be used as database for BLAST search. Therefore, if the query sequences have orthologous proteins in pan-genome set, the queried proteins will be annotated accordingly. Thus, by performing the BLAST search against any user-defined pan-genomic profile for all the proteins in any new genome of interest, it is possible to define the core, accessory and unique proteins of that new genome.

Demonstration of Comparative Analysis and Its Applications

For demonstration of Pan-CA Module, we considered all available *Lactobacillus* strains from human microbiome and divided them into three groups according to their major body-site of isolation, i.e., human gastrointestinal tracts (gut), oral cavity and urogenital tracts. The summary of selected dataset is shown in **Table 1**. The complete list of strains used for this analysis is provided in **Supplementary Table S2**.

These three groups are provided as input for comparative analysis to retrieve group specific exclusive sets of gene families, KEGG enzymes, and COG annotated proteins. The analysis reveals interesting trend about the peculiar gene/function repertoire of these three groups and created a comparative evolutionary portrait of *Lactobacillus* strains at the distinct body-sites.

The Gene Family Distribution

The complete set of proteins upon clustering (using sequence identity cut-off of 50%) generates the protein families for all members of three groups. The group specific exclusive sets are calculated along with all other possible combinations between groups. Then the shared and exclusive gene sets are extracted with sequences. As shown in **Figure 3A**, there are 2477 gene families which contain proteins from at least

one member from each body-site. Out of these 2477 gene families, 68 gene families are found in all the 67 *Lactobacillus* strains irrespective of body-sites representing the *absolute core*; most of them are involved in house-keeping functions like *translation* and *cell wall/membrane/envelop biogenesis*. While, the remaining 2409 gene families represent *extended core* set. There are 10185, 5557 and 2059 gene families specific for gut, urogenital tract and oral cavity respectively (**Figure 3A**).

The COG Function Distribution

The comparison of COG identifiers pooled together for each *Lactobacillus* group provides exclusive COG functions present at respective body-site at annotation level; irrespective of strain details. **Figure 3B** shows the distribution of COGs between the three *Lactobacillus* sets. There are total 1348 COGs common to all the three *Lactobacillus* groups, i.e., core in nature. As shown in **Figure 3F**, most of the Core COGs fall under *Amino acid transport and metabolism, Translation* and *Carbohydrate transport and metabolism*. The distributions of body-site specific functional categories are also retrieved through Pan-CA module (see **Supplementary Figures S2–S4**).

The KEGG Enzyme Distribution

The comparison of KEGG enzymes pooled together for each *Lactobacillus* group provides exclusive pathway profile present at respective body-site at pathway level; irrespective of strain details (**Figure 3C**). Most out of these 1192 core enzymes are involved in *Metabolism* and *Genetic information processing* (**Figure 3D**). Upon detailed analysis, the proportions of genes in *Translation, Carbohydrate metabolism*, and *Membrane transport* pathway categories are found to be high within these core enzymes (**Figure 3E**).

The results also reveal about gut, oral cavity and urogenital tract specific enzymes among *Lactobacillus* strains (see **Supplementary Figures S5–S10**). Overall, the gut, oral and urogenital tract specific enzymes show highest proportion of *Membrane transport* related pathways. However, the *Cell motility* pathways are highly represented in gut specific Lactobacilli; this is in conformation of previous reports suggesting biological significance for presence of cell motility in gut bacteria which may potentially favor better acquisition of nutrients and successful colonization to the niche environment (Cousin et al., 2015). *Cancer* related pathways are present in oral Lactobacilli only, indicating possible role of oral microbiota in carcinogenesis (Meurman, 2010). *Signal transduction* related pathways are in higher proportion in urogenital tract Lactobacilli as compared to those from other body-sites, also reported previously (Mendes-Soares et al., 2014). Such body-site specific enzyme sets might be involved in body-site specific adaptive strategies during human-microbe co-evolution.

Comparison of PanGFR-HM With Other Resources

PanGFR-HM is the only resource providing comprehensive pan-genomic analysis exclusively for the human microbiome strains. Also, as per our knowledge, no resource provides online comparative gene, COG/KEGG classification analysis of user-defined groups of microbiome strains. However, some related resources are considered here for overall comparison of pan-genomic output on microbial data irrespective of their relation to human microbiome context. The details can be accessed from **Table 2**.

DISCUSSION

The prime objective of PanGFR-HM was to create a user friendly dynamic platform, which applies concept of pan-genome to better understand genomic/functional repertoire of inhabitant

TABLE 2 | Comparison of PanGFR-HM with other microbial pan-genome analysis resources.

Functional feature	PanGFR-HM†	EDGAR 2.0	Micro Scope	MetaRef
No. of genomes included	1293 (S.)	2160 (N.S.)	3871 (N.S.)	2818 (N.S.)
Pan-genome distribution	✓	✓	✓	✓
Pan, core profile (development) plots	✓	✓	✓	×
Strain wise pan-genome distribution	✓	✓	✓	×
Core-genome based phylogeny	✓	✓	×	×
Forming groups of strains	✓	✓	×	×
Sequence identity cut-off setting option	✓#	×	✓*	×
Editable and interactive plots	✓	×	×	×
Pan-genome based phylogeny	✓	×	×	×
Strain wise sequence retrieval from pan-genome	✓	×	×	×
Exclusive gene distribution	✓	×	×	×
Exclusive COG and KEGG distribution	✓	×	×	×
Over/under-representation of functional classes	✓	×	×	×
Exclusive absence of genes	✓	×	×	×
COG and KEGG distribution of pan-genome	✓	×	×	×

†*Present tool,* #*40–90% (40, 50, 60, 70, 80, and 90%) identity cut-off options available,* **Only 50 and 80% identity cut-off option available, S., Specific to human microbiome, N.S., Non Specific,* ✓*-present,* ×*-absent. Features listed in italics are exclusive to PanGFR-HM.*

microbes of the human microbiome. This web resource is equipped with unique features to extrapolate the genomic data to speed up and simplify pan-genomic and functional comparative analyses on large datasets of reference microbes from the human body.

Limitations of the Pan-Genome Construction Methods

In cases of orthology based pan-genome approaches, the sequence identity cut-off is the critical parameter which determines if the given gene family belongs to conserved genome or dispensable genome. Larger changes in the cut-off values may considerably change the status of gene family. The higher identity cut-off (more than 70%) may reduce the 'core' set and increase the accessory or strain specific gene sets. On the other hand, lower identity cut-off used for exactly same dataset will allow more genes to be assigned as core genes based on lower threshold for ortholog prediction. Also, the protein diversity within a selected taxon, clade or dataset is one of the factors for deciding appropriate identity cut-off. The members of same species are closely related in taxonomic and evolutionary aspects. They need higher identity cut-offs to establish the orthology in order to reveal recent evolutionary changes. As we move from specific taxonomic levels like species to genus or more general ones, the members become distant in terms of genome evolution, so, lower identity cut offs are recommended. So, the default 50% used for PanGFR-HM seems optimal for related organisms up to genus or family level, but again the genome diversity characteristics of each genus or family may vary. The users need to set these parameters with caution.

Availability of Complete Genomes for Human Microbiota

The present dataset of completely sequenced microbial genomes isolated from human body specific sources may not represent the complete picture of the microbiome, it will always remain a work in progress for a while. The advantage of pan-genome based concept is that it hints you toward the sequencing effort needed for certain taxa, i.e., whether the number of strains used in pan-genome are sufficient to explain the genomic architecture of particular taxon. For taxa showing open pan-genomes need more and more completed genomes of its members for more comprehensive genomic landscape of those taxa, while the near-closed pan-genome suggests limited gene acquisition and loss within that taxon.

CONCLUSION

This resource will encourage researchers to study essential and ubiquitous microbiota at various taxonomic levels and enable them to gaze into the intricate functional and pathway details of specific groups of microbiome communities. Currently, the resource is focused to the genomic/functional repertoire of completely sequenced microbial genomes from HMP, and in future we plan to make the database more resourceful with each update by incorporating new complete genomes, draft genomes and genomes from other sources. As there will be additional newly sequenced complete microbiome stains available through human microbiome or other microbiome projects we plan to update the database contents twice a year to accommodate those strains. Obviously, the more the reference genomes better will be the overall representation of pan-genomic features. PanGFR-HM is committed to accommodate the expanding taxonomic and genomic landscape of the human microbiome.

AVAILABILITY OF SUPPORTING DATA AND MATERIALS

The resource can be freely accessed at http://www.bioinfo.iicb.res.in/pangfr-hm/. All the complete genomes used for generation of PanGFR-HM were publically available from https://www.hmpdacc.org/HMRGD/ (The complete list of genomes used for PanGFR-HM is available in **Supplementary Table S1**). The case study results on 67 *Lactobacillus* strains can be reproduced from http://www.bioinfo.iicb.res.in/pangfr-hm/pan-ca.html, by selecting the strains listed in **Supplementary Table S2**.

AUTHOR CONTRIBUTIONS

NC and VG conceptualized the project and drafted manuscript. NC, VG, AG, and GK generated pan genome database from raw data. NC and AG did the required programming. CD added thoughtful suggestions during the work and manuscript writing. SP conceived and coordinated the project, and revised the manuscript. All the authors read and approved the final manuscript.

FUNDING

NC was supported by Senior Research Fellowship from CSIR, Government of India. SP was supported by Ramanujan Fellowship of Science and Engineering Research Board (SERB), Government of India. The work was supported by the Ramanujan Fellowship Grant of SERB and Systems Medicine Cluster (SyMeC) Project, Department of Biotechnology, Government of India.

ACKNOWLEDGMENTS

We would like to acknowledge Dr. Sucheta Tripathy for technical support and server provisions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02322/full#supplementary-material

REFERENCES

Arber, W. (2014). Horizontal gene transfer among bacteria and its role in biological evolution. *Life* 4, 217–224. doi: 10.3390/life4020217

Bakshi, U., Sarkar, M., Paul, S., and Dutta, C. (2016). Assessment of virulence potential of uncharacterized enterococcus faecalis strains using pan genomic approach - identification of pathogen-specific and habitat-specific genes. *Sci. Rep.* 6:38648. doi: 10.1038/srep38648

Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C., et al. (2016). EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Res.* 44, W22–W28. doi: 10.1093/nar/gkw255

Bolotin, E., and Hershberg, R. (2015). Gene loss dominates as a source of genetic variation within clonal pathogenic bacterial species. *Genome Biol. Evol.* 7, 2173–2187. doi: 10.1093/gbe/evv135

Chaudhari, N. M., Gupta, V. K., and Dutta, C. (2016). BPGA- an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* 6:24373. doi: 10.1038/srep 24373

Conlan, S., Mijares, L. A., NISC Comparative Sequencing Program. Becker, J., Blakesley, R. W., Bouffard, G. G., et al. (2012). Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. *Genome Biol.* 13:R64. doi: 10.1186/gb-2012-13-7-r64

Cousin, F. J., Lynch, S. M., Harris, H. M. B., McCann, A., Lynch, D. B., Neville, B. A., et al. (2015). Detection and genomic characterization of motility in *Lactobacillus curvatus*: confirmation of motility in a species outside the *Lactobacillus salivarius* clade. *Appl. Environ. Microbiol.* 81, 1297–1308. doi: 10.1128/AEM.03594-14

Duranti, S., Milani, C., Lugli, G. A., Mancabelli, L., Turroni, F., Ferrario, C., et al. (2016). Evaluation of genetic diversity among strains of the human gut commensal *Bifidobacterium adolescentis*. *Sci. Rep.* 6:23971. doi: 10.1038/srep23971

Dutta, C., and Pan, A. (2002). Horizontal gene transfer and bacterial diversity. *J. Biosci.* 27, 27–33. doi: 10.1007/BF02703681

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

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

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* 43, D261–D269. doi: 10.1093/nar/gku1223

Gupta, V. K., Chaudhari, N. M., Iskepalli, S., and Dutta, C. (2015). Divergences in gene repertoire among the reference *Prevotella genomes* derived from distinct body sites of human. *BMC Genomics* 16:153. doi: 10.1186/s12864-015-1350-6

Gupta, V. K., Paul, S., and Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front. Microbiol.* 8:1162. doi: 10.3389/fmicb.2017.01162

Hottes, A. K., Freddolino, P. L., Khare, A., Donnell, Z. N., Liu, J. C., and Tavazoie, S. (2013). Bacterial adaptation through loss of function. *PLoS Genet.* 9:e1003617. doi: 10.1371/journal.pgen.1003617

Huang, K., Brady, A., Mahurkar, A., White, O., Gevers, D., Huttenhower, C., et al. (2014). Metaref: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Res.* 42, D617–D624. doi: 10.1093/nar/gkt1078

Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi: 10.1093/nar/gkw290

Li, S. J., Hua, Z. S., Huang, L. N., Li, J., Shi, S. H., Chen, L. X., et al. (2014). Microbial communities evolve faster in extreme environments. *Sci. Rep.* 4:6205. doi: 10.1038/srep06205

Mendes-Soares, H., Suzuki, H., Hickey, R. J., and Forney, L. J. (2014). Comparative functional genomics of *Lactobacillus spp*. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *J. Bacteriol.* 196, 1458–1470. doi: 10.1128/JB.01439-13

Meurman, J. H. (2010). Oral microbiota and cancer. *J. Oral. Microbiol.* 2:5195. doi: 10.3402/jom.v2i0.5195

Moriya, Y., Itoh, M., Okud;a, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321

NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., et al. (2009). The NIH human microbiome project. *Genome Res.* 19, 2317–2323. doi: 10.1101/gr.096651.109

Paul, S., Minnick, M. F., and Chattopadhyay, S. (2016). Mutation-driven divergence and convergence indicate adaptive evolution of the intracellular human-restricted pathogen, *Bartonella bacilliformis*. *PLoS Negl. Trop. Dis.* 10:e0004712. doi: 10.1371/journal.pntd.0004712

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinformatics* 42, 3.1.1–3.1.8. doi: 10.1002/0471250953.bi0301s42

Popa, O., Hazkani-Covo, E., Landan, G., Martin, W., and Dagan, T. (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* 21, 599–609. doi: 10.1101/gr.115592.110

Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., et al. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* 190, 6881–6893. doi: 10.1128/JB.00619-08

Reid, G., Younes, J. A., Van der Mei, H. C., Gloor, G. B., Knight, R., and Busscher, H. J. (2011). Microbiota restoration: natural and supplemented recovery of human microbial communities. *Nat. Rev. Microbiol.* 9, 27–38. doi: 10.1038/nrmicro2473

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Vallenet, D., Calteau, A., Cruveiller, S., Gachet, M., Lajus, A., Josso, A., et al. (2017). Microscope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Res.* 45, D517–D528. doi: 10.1093/nar/gkw1101

Conflict of Interest Statement: 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.

Copyright © 2018 Chaudhari, Gautam, Gupta, Kaur, Dutta and Paul. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 09 October 2018
doi: 10.3389/fmicb.2018.02440

Human Microbe-Disease Association Prediction Based on Adaptive Boosting

Li-Hong Peng[1†], Jun Yin[2†], Liqian Zhou[1], Ming-Xi Liu[3] and Yan Zhao[2*]*

[1] School of Computer Science, Hunan University of Technology, Zhuzhou, China, [2] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China, [3] Institutes of Science and Development, Chinese Academy of Sciences, Beijing, China

There are countless microbes in the human body, and they play various roles in the physiological process. There is growing evidence that microbes are closely associated with human diseases. Researching disease-related microbes helps us understand the mechanisms of diseases and provides new strategies for diseases diagnosis and treatment. Many computational models have been proposed to predict disease-related microbes, in this paper, we developed a model of Adaptive Boosting for Human Microbe-Disease Association prediction (ABHMDA) to reveal the associations between diseases and microbes by calculating the relation probability of disease-microbe pair using a strong classifier. Our model could be applied to new diseases without any known related microbes. In order to assess the prediction power of the model, global and local leave-one-out cross validation (LOOCV) were implemented. As shown in the results, the global and local LOOCV values reached 0.8869 and 0.7910, respectively. What's more, 10, 10, and 8 out of the top 10 microbes predicted to be most likely to be associated with Asthma, Colorectal carcinoma and Type 1 diabetes were all verified by relevant literatures or database HMDAD, respectively. The above results verify the superior predictive performance of ABHMDA.

Keywords: microbe, disease, association prediction, adaptive boosting, decision tree

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Qinghua Cui,
Peking University, China

***Correspondence:**
Ming-Xi Liu
liumingxi@casipm.ac.cn
Yan Zhao
ts17060090a3@cumt.edu.cn

†Joint first authors

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

Received: *19 August 2018*
Accepted: *24 September 2018*
Published: *09 October 2018*

Citation:
Peng L-H, Yin J, Zhou L, Liu M-X and Zhao Y (2018) Human Microbe-Disease Association Prediction Based on Adaptive Boosting. Front. Microbiol. 9:2440. doi: 10.3389/fmicb.2018.02440

INTRODUCTION

Microbes are ubiquitous in our lives. After deeper research, microbes could be simply divided into the following types: bacteria, fungi, viruses, archaea, protozoa, and so on (Sommer and Backhed, 2013). As we all know, there are a number of microbes living in the human tissues, such as gut (Grenham et al., 2011), skin (Fredricks, 2001) and lung (Cole, 1989). Cells are the basic unit of our body's structure and function, and our body contains more than 40 trillion cells, but studies have shown that the number of microorganisms in humans is 10% more than the number of cells, which shows that the microbial community is relatively large in the human body (Sender et al., 2016). There are studies showing that microorganisms are involved in many biological processes in the human body, such as metabolic function, immune function, and so on (Gill et al., 2006). For example, in the intestinal tract of the adult, most of the intestinal microbes living in the gastrointestinal tract are able to not only synthesize necessary amino acids and vitamins, but also are conducive to the digestion and absorption of indigestible food (Huang Z.A. et al., 2017). So it is not surprising that there are links between microbes and human diseases (Consortium, 2012).Some researchers had found a close relationship between human type 2 diabetes and changes in the composition of the intestinal microbiota (Larsen et al., 2010). Gut microbes could induce colorectal cancer by generating butyrate that

promoted the hyperproliferation of MSH2(−/−) colon epithelial cells (Belcheva et al., 2014). There was also evidence that toxins produced by microbes such as Streptococcus and Staphylococcus aureus had been shown to be a new class of allergens that could induce or even aggravate inflammatory skin diseases (Skov and Baadsgaard, 2000). Therefore, revealing disease-related microbes not only helps to further understand the pathogenesis of the disease but also provides new strategies for the diagnosis and treatment of the disease. Although some proven disease-microbe associations have been documented in the database HMDAD (Ma et al., 2017)[1], such as Allergic asthma-Helicobacter pylori, Allergic sensitization-Clostridium difficile, and Asthma-Bacteroidetes, these are far from enough. Unfortunately, using biological experiments to reveal the relationship between disease and microbes is cumbersome and costly. Therefore, it is imperative to predict the potential disease-related microbes by constructing computational models.

According to the assumption that functionally similar microbes tend to be associated with similar diseases, by integrating two separate recommendation algorithms based on neighbor information and network topology, respectively, Huang Y.A. et al. (2017) developed a neighbor and graph based combined recommendation model for human microbe-disease association prediction (NGRHMDA) to predict potential disease-related microbes. As a combination of two independent recommendation models, the prediction accuracy of NGRHMDA was significantly improved compared to a single recommendation model. Unlike previous methods, NGRHMDA was an unsupervised learning method that did not require negative samples. Of course, there were some restrictions on NGRHMDA. Firstly, NGRHMDA could not be applied to predict microbes associated with new diseases without any known related microbes. Secondly, the optimal values of some parameters in the model were still not solved. Huang Z.A. et al. (2017) proposed a method of Path-Based Human Microbe-Disease Association prediction (PBHMDA) by integrating confirmed disease-microbe relations and the Gaussian interaction profile kernel similarity for diseases and microbes into a heterogeneous network. This model traversed all possible pathways between microbes and diseases through a novel depth-first search algorithm to predict the most likely disease-associated microbes. Both global and local leave-one-out cross validation (LOOCV) AUC values of PBHMDA were greater than 0.9, which showed that the prediction accuracy of PBHMDA was quite impressive. Regrettably, this model still had some shortcomings. Firstly, both the disease–disease similarities and microbe–microbe similarities were obtained from the Gaussian kernel for interaction profiles of microbes and diseases that were calculated based on the known disease-microbe associations, which might be biased for diseases with more known related microbes. Secondly, PBHMDA was also not suitable for new diseases. What's more, based on the known human microbe-disease association network obtained from the HMDAD database, Wang et al. (2017) proposed a novel computational model of Laplacian Regularized Least Squares for Human Microbe-Disease Association (LRLSHMDA) to reveal potential disease-related microbes (Wang et al., 2017). LRLSHMDA applied a semi-supervised learning framework due to the lack of pairs of disease-microbes that had proven to be unrelated. In this model, the microbe similarity network and the disease similarity network were constructed based on the Gaussian interaction profile kernel similarity calculated by known microbe-disease association, and then by constructing and optimizing the cost functions in microbe space and disease space to integrated the optimal classifier functions to calculate the relation probabilities of microbe-disease pairs. Although the reliable prediction performance of LRLSHMDA had been verified, the model still had some shortcomings that needed further improvement. Firstly, the number of proven-microbe associations was too small, and sparse known association network might affect the prediction performance of the model. Secondly, LRLSHMDA could not be suitable for new microbes without any known related diseases.

In addition, Ma et al. (2017) built a microbe-disease association network based on published literature, and constructed a disease–disease network (Human Microbe Disease Network, HMDN) based on disease-associated microbes where the weight of the link between diseases was the similarity of microbes associated with the corresponding disease, and then by integrating data of disease genes, symptoms, chemical fragments, and drugs to investigate the overlaps between microbes and genes. Chen et al. (2017a) built a microbe-human disease association network and proposed a novel computational model of KATZ measure for Human Microbe-Disease Association prediction (KATZHMDA) based on this hypothesis that functionally similar microbes tend to have similar interactions and non-interactive patterns with non-infectious diseases and vice versa. By merging known disease-microbe association networks, disease similarity networks and microbe similarity networks into a heterogeneous network, KATZHMDA integrated walks with different lengths in the network to calculate the relation probability between microbe and disease. As a global computation method, KATZHMDA was capable of simultaneously revealing microbes associated with all diseases in a large-scale network. However, KATZHMDA still had many problems need to be solved in the future. For example, the problem of the optimal value of the parameter k had not been solved yet, and the prediction accuracy of KATZHMDA needed to be improved.

The above methods had various shortcomings. For instance, some models were not suitable for new diseases, and the optimal values of the parameters in some models were not well solved. For the sake of revealing the association between microbe-diseases better, in this paper, we proposed a model of Adaptive Boosting for Human Microbe-Disease Association prediction (ABHMDA) to uncover the associations between diseases and microbes by calculating the relation probability of disease-microbe pair using a strong classifier. Compared with the above methods, our model had the advantage of predicting microbes associated with new diseases. Since the number of negative samples was much larger than that of positive samples, we introduced k-means clusters to sample negative samples to balance the samples for training.

[1]http://www.cuilab.cn/hmdad

What's more, the strong classifier was composed of multiple weak classifiers according to the corresponding weights, and the higher the prediction accuracy of weak classifier, the greater the weight of it. We applied global and local LOOCV to evaluate the prediction performance of ABHMDA. As the results shown, the global and local LOOCV values reached 0.8869 and 0.7910, respectively, which indicated that the model's prediction power was reliable. Besides, we used ABHMDA to conduct case studies on three diseases. 10, 10, and 8 out of the top 10 microbes predicted to be most likely to be associated with Asthma, Colorectal carcinoma and Type 1 diabetes were all verified by relevant literatures or database HMDAD, respectively.

MATERIALS AND METHODS

Human Microbe-Disease Associations

We could obtain 450 known associations between 292 microbes and 38 diseases from Human Microbe-Disease Association Database (HMDAD) (Ma et al., 2017). For the reason that there were several grades of microbe classification, and when using 16s RNA sequences to study microbes, only the information in the level of genus would be acquired, we revealed the microbes which were likely to be related with human diseases in genus level. Besides, we defined the adjacency matrix A, if there was known association between disease $d(i)$ and microbes $m(j)$, the value of the element $A(d(i), m(j))$ matrix A was 1. We applied the variable nd, nm to denote the number of diseases and microbes studied, respectively.

Gaussian Interaction Profile Kernel Similarity

Inspired by this article (Laarhoven et al., 2011), Considering the assumption that if two similar diseases were associated with two microbes, respectively, the two microbes were likely to be similar, and there were similar interaction and non-interaction pattern between diseases and microbes, Gaussian interaction profile kernel similarity for disease KD was constructed to indicated the similarities between diseases based on the known associations of disease-microbe pairs. Firstly, binary vector $IP(d(i))$was defined to represented the interaction profiles of diseases $d(i)$ by observing whether there was a known association between disease $d(i)$ and each microbe (i.e., the ith row of the adjacency matrix A). Then, the Gaussian interaction profile kernel similarity between disease $d(i)$ and $d(i)$ could be calculated as follow:

$$KD(d(i), d(j)) = \exp\left(-\gamma_d ||IP(d(i)) - IP(d(i))||^2\right) \quad (1)$$

Here, parameter γ_d was introduced to regulated the kernel bandwidth and got by normalizing another parameter γ'_d by the average number of related microbes of all the diseases. γ_d was calculated as follow:

$$\gamma_d = \frac{\gamma'_d}{\frac{\sum_1 nd||IP(d(i))||^2}{nd}} \quad (2)$$

where the value of γ'_d was 1.

The definition of Gaussian interaction profile kernel similarity for microbe KM was similar to KD

Integrating Symptom-Based Disease Similarity

From the above we could see that Gaussian interaction profile kernel similarity was only based on the adjacency matrix A. If we wanted to effectively and scientifically predict potential disease-associated microbes, it was necessary to introduce other datasets in combination with the Gaussian interaction profile kernel similarity. Based on the disease and corresponding symptom recorded in PubMed bibliography. Zhou et al. (2014) calculated similarity between diseases and constructed the symptom-based human disease network (HSDN). Here, we integrated the Gaussian interaction profile kernel similarity for disease KD and the symptom-based disease similarity SDM to obtained the Integrating symptom-based disease similarity SD, and the calculation of SD was defined as follow:

$$SD = \frac{KD + SDM}{2} \quad (3)$$

ABHMDA

Motivated by this paper (Rayhan et al., 2017), we constructed a novel calculation model of ABHMDA to predict disease-related microbes and the flow chart of the algorithm was shown in **Figure 1**. The core idea of ABHMDA was to train different classifiers (weak classifiers) for the same training samples, and then grouped these weak classifiers with different ratios to form a stronger classifier to score and sort samples. Here, we chose the decision tree as our weak classifier. The specific steps were mainly divided into three steps: integrating the data, training the model, and scoring the samples. In the first step, we integrated the Gaussian interaction profile kernel similarity for microbe KM and the Integrating symptom-based disease similarity SD. In the second step, we firstly referred to the sample with confirmed association as a positive sample, otherwise it was an unknown sample. On account of the unknown sample accounting for about 97% in all the samples, that was to say, there were far more unknown samples than positive ones, and it was unreasonable to directly train such unbalanced datasets. Here, we introduced a novel method to balance the datasets. In this method, we applied the k-mean clustering to divide the unknown sample into k parts, and then randomly extract some samples from each part as negative samples, while positive samples kept unchanged. There were researchers studying the effect to random extraction when k took different values, and the results shown that the optimal value of parameter k was 23. In order to make the dataset used for training more balanced, the number of the unknown samples randomly selected ought to be approximately equal to the positive sample. In the end, the negative and positive samples together formed the training samples. Each training sample was weighted with an initial weight of $\frac{1}{n}$, where n was the total number of training samples. The main purpose of the training process was to calculate the proportion of each weak classifier in the final strong classifier and update the weight of each training sample according to whether it was classified correctly by the last classifier and the

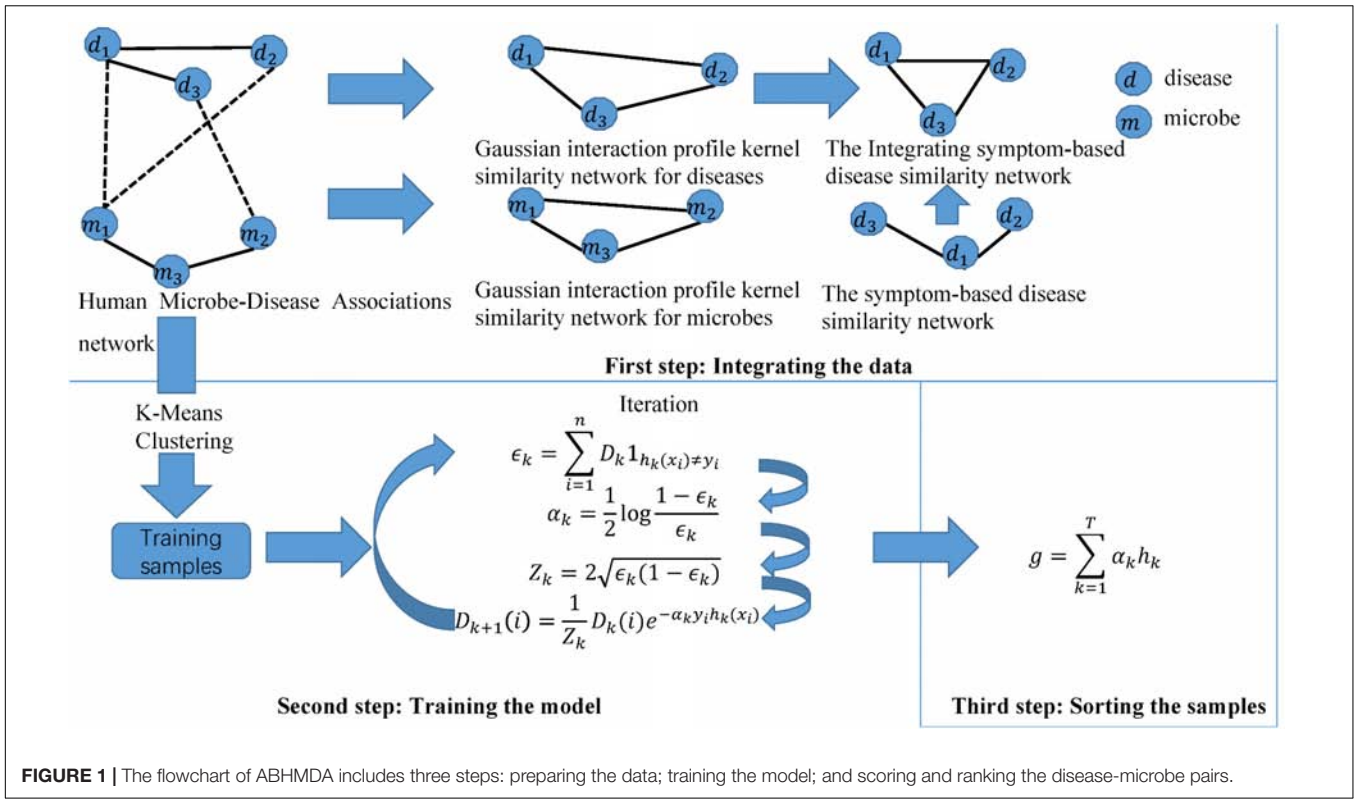


FIGURE 1 | The flowchart of ABHMDA includes three steps: preparing the data; training the model; and scoring and ranking the disease-microbe pairs.

overall classification accuracy of the last classifier. After updating, the new training sample set with modified weight values was sent to the next weak classifier for training. Here, we built lists D_I, $h(i)$ and Y, all of which had n elements. The value of each element in D_{i} was the weight of the corresponding sample when the *ith* weak classifier trained the sample. The value of i was 0, 1, 2, , , , 29. In other words, D_0 was a list with all elements being $\frac{1}{n}$. The value of the element in label lists $h\,(i)$ and Y was only 0 or 1, and the difference between them was that the value of $h(i)_{\mathrm{j}}$ depended on the prediction of the *ith* weak classifier, while the value of Y_{j} depended on whether the corresponding sample was a positive sample, if the corresponding sample was a positive sample, the value of Y_{j} was equal to 1, otherwise 0. The error function $\in_{\mathrm{i}}$ was calculated as follow:

$$\in_{\mathrm{i}} = \sum_{\mathrm{j}=1}^{\mathrm{n}} D_{\mathrm{i}} 1_{h(i)j\neq Y_{\mathrm{j}}} \tag{4}$$

It could be seen from the formula that the error function $\in_{\mathrm{i}}$ was equal to the sum of the weights of the samples, whose label predicted by the weak classifier $h\,(i)_{\mathrm{j}}$ was different from the known label Y_{j}. That was to say $\in_{\mathrm{i}}$ was equal to the sum of the weights of all the samples that were predicted wrong. Then the proportion of the *ith* weak classifier in the strong classifier could be defined as follow:

$$\alpha_{\mathrm{i}} = \frac{\log\frac{1-\in_{\mathrm{i}}}{\in_{\mathrm{i}}}}{2} \tag{5}$$

It could be seen from equation (5) that the smaller the error function was, the larger the proportion of the weak classifier in the strong classifier would be. And the variate Z_{i} could be calculated as follow:

$$Z_{\mathrm{i}} = 2\,[\in_{\mathrm{i}}\,(1-\in_{\mathrm{i}})]^2 \tag{6}$$

The weight of the sample could be updated according to the following formula:

$$D_{\mathrm{i}+1}\,(j) = \frac{1}{Z_{\mathrm{i}}} D_{\mathrm{i}}\,(j)\, e^{-\alpha_{\mathrm{i}} Y_{\mathrm{j}} h(i)_{\mathrm{j}}} \tag{7}$$

Here $j = 0, 1, 2 \ldots n-1$. After the weights of samples being updated, the samples with the new weights were sent to the next weak classifier to start the next training until all the weak classifiers completed the training (Theoretically, the more weak classifiers, the higher the prediction accuracy of strong classifier. But when the weak classifier reached a certain number, the prediction accuracy tended to be stable. And then as the number of weak classifiers increased, accuracy was not significantly improved, but the prediction process took longer. We compared the prediction results with 20, 30, and 40 weak classifiers, the accuracy of using 30 and 40 weak classifiers was basically the same, which was better than 20 weak classifiers. However, the prediction time of 40 weak classifiers was longer than using 30 classifiers. Comprehensive consideration of prediction time and accuracy, here, we chose to use 30 weak classifiers to form the final strong classifier.), then the training process was end. The next step was to score the sample, and the score of the *jth* sample

was defined as follows:

$$s\left(j\right) = \sum_{i=0}^{29} \alpha_i H\left(i\right)_j \quad (8)$$

Here, $H\left(i\right)_j$ was the score scored by the *ith* weak classifier for the *jth* sample. That was to say, the score of the sample was equal to the sum of the product of the sample's goal scored by weak classifier and the corresponding weight (The corresponding data and code had been submitted to the website[2]).

RESULTS

Performance Evaluation

In order to verify the prediction performance of ABHMDA, we implemented global and local LOOCV for our model based on the database HMDAD (Ma et al., 2017) which recorded 450 known associations between 39 diseases and 292 miRNAs. Specifically, each of the 450 samples (positive samples) with known association was left out in turn as a test sample while the remaining 449 were used for model training, while all of the samples without known associations were considered as candidate samples (unknown samples). In global LOOCV, we sorted the test sample with all candidate samples based on the score marked by calculation model, while the test sample was ranked with the candidate samples that contained the same disease as the test sample in local LOOCV. We evaluated the prediction performance of models based on the AUC value of the LOOCV. To be specific, only the test sample ranked above a certain threshold, could it be considered as a correct prediction, and then we set the true positive rate (TPR, sensitivity) as the horizontal axis and the false positive rate (FPR, 1-specificity) as the vertical axis. Therefore, we could plot the Receiver operating characteristics (ROC) curve, which was composed of points corresponding to different thresholds, then we could obtain the Area under the ROC curve (AUC). A model with an AUC value equal to 0.5 was equivalent to a random prediction. When the AUC took the maximum value of 1, the model had excellent prediction performance. In other words, when the value of AUC was greater than 0.5 and less than 1, the larger the value was, the better the prediction performance of the model would be.

As shown in **Figure 2**, the global LOOCV value of ABHMDA was 0.8869, which was significantly larger than that of KATZHMDA (0.8644) and LRLSHMDA (0.8843). What was more, the local LOOCV value of our model reached 0.7910, which was also obviously better than KATZHMDA (0.6998) and LRLSHMDA (0.7508). These results confirmed the superior prediction performance of ABHMDA

Case Study

In order to further assess the prediction ability of ABHMDA, we implemented two case studies on some important diseases of human. In the first kind, there were 10938 unknown samples about 39 diseases and 292 miRNAs in HMDAD. We sorted and ranked all unknown samples corresponding to the same disease and verified whether the association between the top 10 microbes and the disease studied was verified by the relevant literature. In the second kind, we converted all 1 in the adjacency matrix *A* to 0 and sorted all the samples (positive and unknown samples) corresponding to the same disease and then verified the association between disease and the 10 microbes most likely associated with it predicted by the model in the database HMDAD. In other words, the purpose of the second case study was to verify our model's power to predict microbes associated with new diseases without any known related microbes. Here, we implemented the first case study on asthma, Colorectal carcinoma, and the second case on Type 1 diabetes.

As an inflammatory disease on the airway, it was very difficult to completely cure asthma under current medical conditions (Preston et al., 2007). According to statistics, there were about 300 million asthma patients worldwide, and in recent years its morbidity and mortality had also increased rapidly, especially in developing countries (Sagar et al., 2014). Therefore, a deeper study of asthma was imperative, and studies had shown that there was a close relationship between the microbes in the respiratory tract and the development and progression of asthma (Marri et al., 2013). For example, studies had shown that Firmicutes was reduced in asthmatic patients compared with normal humans (Wu et al., 2018). In contrast, Proteobacteria accounted for a larger proportion of microorganisms in asthma patients than normal people (Marri et al., 2013). What's more, there was evidence that when the hypopharyngeal area of Neonates was infected with Streptococcus pneumoniae, the risk of developing asthma was increased compared to uninfected (Bisgaard et al., 2007). We implemented the first case study of asthma and the 10 microbes predicted to be most relevant to asthma were all verified by literatures. For instance, the experimental results showed that the abundance of Lachnospiraceae (First in the prediction list) in asthma patients was 1.9 times that of normal people (Jung et al., 2016). The researchers found that the relative abundance of Veillonella (Second in prediction list) in infants at risk of asthma was significantly lower than in normal people, and inoculation of sterile mice with Veillonella could improve its airway inflammation, which provided new ideas for the treatment of asthma (Arrieta et al., 2015). Moreover, there was evidence that if there was Clostridium coccoides (Third in prediction list) in a 3 week old baby's stool, he was at risk of developing asthma, so Clostridium coccoides may become an early diagnostic target for asthma (Vael et al., 2011; See **Table 1**).

To facilitate further research and validation, we provided a ranking of the relevant probabilities for all pairs of disease-microbe pairs without confirmed association (See **Supplementary Table S1**).

Colorectal carcinoma (CRC) was a common gastrointestinal malignant tumor in China (Xue et al., 2014). As one of the top cancers with the highest morbidity and mortality worldwide, it was estimated that there were approximately one million new cases of CRC and 500000 deaths per year (Sun et al., 2013). What was more serious was that its incidence would continue to increase in the next few decades, and the survival rate in 5 years was less than 60% (Sun et al., 2011). Therefore, it

[2]https://github.com/githubcode007/ABHMDA

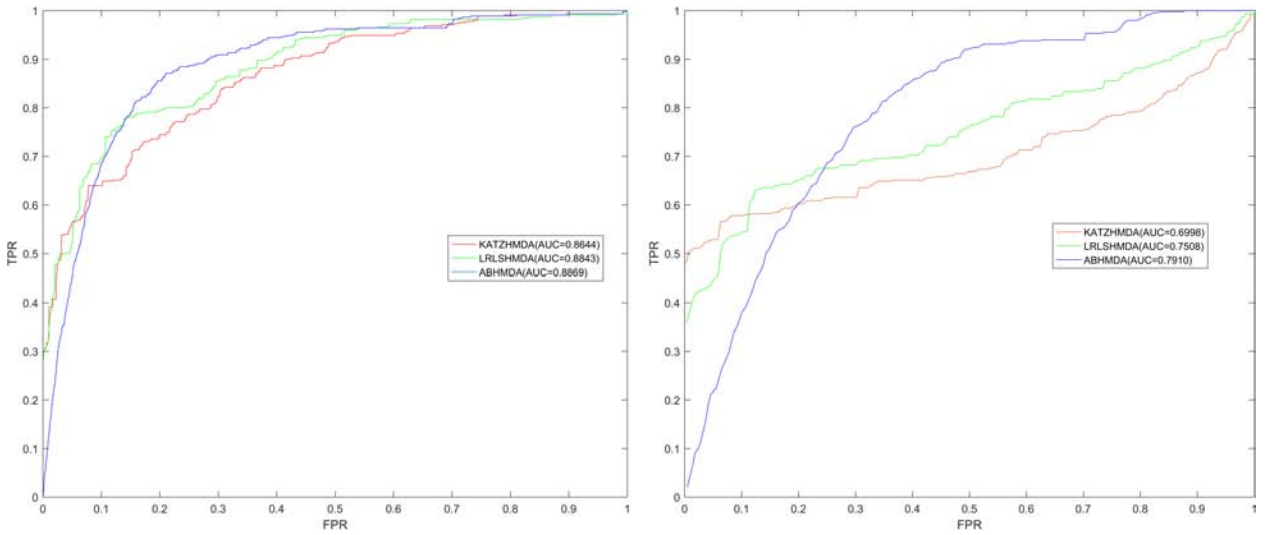


FIGURE 2 | Comparison of prediction performance of ABHMDA with two other computational models (KATZHMDA, LRLSHMDA) in terms of ROC curves and AUCs values based on global and local LOOCV. As shown in the results, the global and local LOOCV values of ABHMDA were 0.8869 and 0.7910, respectively, which were significantly larger than that of KATZHMDA (0.8644, 0.6998) and LRLSHMDA (0.8843, 0.7508).

TABLE 1 | The 10 microbes predicted to be most likely to be associated with the Asthma.

microbe	Evidece
Lachnospiraceae	PMID: 27433177
Veillonella	PMID: 26424567
Clostridium coccoides	PMID: 21477358
Firmicutes	PMID: 23265859
Streptococcus	PMID: 17928596
Actinobacteria	PMID: 23265859
Lactobacillus	PMID: 20592920
Bacteroides uniformis	PMID: 27433177
Enterococcus	PMID: 22641478
Escherichia coli	PMID: 26277095

The first column records the top 10 microbes most likely to be related Asthma, and the second column records the databases and experimental literatures in PubMed, which verify the associations between the corresponding microbe and Asthma.

TABLE 2 | The 10 microbes predicted to be most likely to be associated with the Colorectal carcinoma.

microbe	Evidece
Veillonella	PMID: 22761885
Klebsiella	PMID: 22776247
Enterobacteriaceae	PMID: 25182170
Proteobacteria	PMID: 24603888
Lachnospiraceae	PMID: 21850056
Clostridium coccoides	PMID: 19807912
Streptococcus	PMID: 21247505
Actinobacteria	PMID: 24316595
Lactobacillus	PMID: 15828052
Bacteroides uniformis	PMID: 24828543

The first column records the top 10 microbes most likely to be related Colorectal carcinoma, and the second column records the databases and experimental literatures in PubMed, which verify the associations between the corresponding microbe and Colorectal carcinoma.

was necessary to study the pathogenesis of CRC to explored new treatment methods, and studies had shown that microbes played an important role in the development and progression of cancer that were closely related to inflammation like CRC (Liang et al., 2014). For example, there were studies showing that the number of Lactobacillus hamster increased significantly during the formation of CRC (Liang et al., 2014). The researchers compared CRC cases with the normal control group and found that the relative abundance of phylum Bacteroidetes in the case group reached 16.2%, which was much higher than 9.9% of the normal group (Ahn et al., 2013). We applied ABHMDA to implement the first case study on CRC, and the 10 predicted microorganisms most likely to be associated with CRC were all verified by related literature in PubMed. There was evidence that the relative abundance of Veillonella (First in the prediction list) in CRC cancer tissues was 2.87% and only 0.68% in the intestinal lumen (Chen et al., 2012). Pyogenic liver abscess was identified as an early manifestation of adult CRC, and an 11-year follow-up study showed that pyogenic liver abscess patients with Klebsiella (Second in the prediction list) pneumoniae had a higher probability of having CRC than those without (Huang et al., 2012). What was more, there were studies showing that Enterobacteriaceae (Third in the prediction list) was very rich in CRC patients (Arthur et al., 2014). From the above results, it could be seen that the predicted performance of ABHMDA was very reliable (See **Table 2**).

Type 1 diabetes was an autoimmune disease which resulted from the immune-mediated destruction of insulin-producing pancreatic β cells (Li et al., 2014). The incidence of Type 1 diabetes was increasing globally, but the proportion of patients

TABLE 3 | The 10 microbes predicted to be most likely to be associated with the Type 1 diabetes.

microbe	Evidece
Veillonella	confirmed
Bacteroidaceae	confirmed
Enterobacteriaceae	PMID: 24475780
Coxiellaceae	unconfirmed
Prevotella	confirmed
Bacteroidetes	confirmed
Prevotella copri	unconfirmed
Lachnospiraceae	confirmed
Lactobacillus	confirmed
Clostridia	confirmed

The first column records the top 10 microbes most likely to be related Type 1 diabetes, and the second column records the databases and experimental literatures in PubMed, which verify the associations between the corresponding microbe and Type 1 diabetes.

suffering from genetic factors was decreasing, which suggested that the virus, nutrition, and overweight were very likely to have become the main cause of Type 1 diabetes (Islam et al., 2014). Studies had shown that the abnormality in the gut microbiota was closely related to the development of Type 1 diabetes (De Goffau et al., 2014). The number of Firmicutes and Actinomycetes were significantly reduced in children with Type 1 diabetes compared with normal people (Murri et al., 2013). We conducted the second case study on Type 1 diabetes to test the prediction power of ABHMDA to predict the potential microbe-related of new diseases, and the results showed that 7 of the top 10 potential disease-related microbes predicted were validated by the database HMDAD. The associations between Type 1 diabetes and microbe Veillonella (First in the prediction list) with Bacteroidaceae (Second in the prediction list) were confirmed by HANDAD. Some researchers had found that patients with Type 1 diabetes had increased colonization of Enterobacteriaceae (Third in the prediction list) in addition to Escherichia coli compared with normal people (Soyucen et al., 2014). The above results indicated that ABHMDA's ability to predict microbes associated with new diseases was also reliable (See **Table 3**).

DISCUSSION

As a kind of tiny creature that are invisible to the human eyes, the microbes are small in size and simple in structure, but they are closely related to human beings. There are thousands of microbes in the human body. They build complex functional institutions and play an extremely important role in many biological processes, although they can benefit people, they can also bring a lot of trouble to human beings, such as diseases. More and more research shows that many human diseases are closely related to microorganisms, especially gastrointestinal diseases. Revealing the relation between disease and microbes contributes to further understand the pathogenesis of the disease and the development of new drugs (Chen et al., 2016b, 2017a). However, due to limited technology, the cost of using experimental methods to reveal disease-related microbes is greater. Therefore, it is imperative to construct model for the prediction of potentially relevant microbes. In this paper, we proposed a novel model ABHMDA to reveal the association between disease and microbes. The global and local LOOCV value of ABHMDA was 0.8869 and 0.7910, respectively, which was significantly larger than that of KATZHMDA (0.8644, 0.6998) and LRLSHMDA (0.8843, 0.7508). This result confirmed the strong prediction power of ABHMDA.

Several factors that led to ABHMDA prediction performance were summarized as follows. Firstly, the datasets used by our model were relatively reliable. Secondly, we extracted the potential similarities for diseases and microbes through Gaussian interaction profile kernel similarity. Thirdly, we combined multiple weak classifiers into one strong classifier according to different weights to score the samples. The high-precision weak classifiers accounted for a high proportion and vice versa, which conduced to improve the accuracy of the strong classifier. Of course, ABHMDA also had some defects that needed to be resolved in future work. Firstly, although the prediction performance of ABHMDA had improved compared to previous methods, prediction capabilities were expected to improve further if more reliable similarities were considered. Many groups have developed several effective computational models for the association prediction (Chen and Yan, 2013; Chen et al., 2016a; Chen and Huang, 2017; You et al., 2017; Chen et al., 2018a,b,c). We would introduce these reliable techniques to this new research area. Secondly, ABHMDA might cause bias to microbes with more associated diseases. Finally, the model did not consider the microbe–microbe similarity based on sequence similarity, which was also where we needed to improve in our future work (Chen et al., 2017b,c; Hu et al., 2018; Zhao et al., 2018).

AUTHOR CONTRIBUTIONS

L-HP and JY implemented the experiments, analyzed the result, and wrote the paper. LZ and M-XL analyzed the result and revised the paper. YZ conceived the project, developed the prediction method, designed the experiments, analyzed the result, and revised the paper. All authors read and approved the final manuscript.

FUNDING

JY and YZ was supported by National Natural Science Foundation of China under grant no. 61772531. L-HP was supported by 61803151 and Natural Science Foundation of Hunan province under grant no. 2018JJ3570. M-XL was supported by the China Postdoctoral Science Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02440/full#supplementary-material

REFERENCES

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

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

Arthur, J. C., Gharaibeh, R. Z., Muhlbauer, M., Perez-Chanona, E., Uronis, J. M., Mccafferty, J., et al. (2014). Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer. *Nat. Commun.* 5:4724. doi: 10.1038/ncomms5724

Belcheva, A., Irrazabal, T., Robertson, S. J., Streutker, C., Maughan, H., Rubino, S., et al. (2014). Gut microbial metabolism drives transformation of MSH2-deficient colon epithelial cells. *Cell* 158, 288–299. doi: 10.1016/j.cell.2014.04.051

Bisgaard, H., Hermansen, M. N., Buchvald, F., Loland, L., Halkjaer, L. B., Bonnelykke, K., et al. (2007). Childhood asthma after bacterial colonization of the airway in neonates. *N. Engl. J. Med.* 357, 1487–1495. doi: 10.1056/NEJMoa052632

Chen, W., Liu, F., Ling, Z., Tong, X., and Xiang, C. (2012). Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS One* 7:e39743. doi: 10.1371/journal.pone.0039743

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

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

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016a). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput. Biol.* 12:e1004975. doi: 10.1371/journal.pcbi.1004975

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2017b). MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* doi: 10.1093/bib/bbx130 [Epub ahead of print].

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017c). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016b). Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018a). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* doi: 10.1093/bioinformatics/bty503 [Epub ahead of print].

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018b). BNPMDA: bipartite network projection for mirna-disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Cole, P. (1989). Host-microbe relationships in chronic respiratory infection. *Respiration* 55(Suppl. 1), 5–8. doi: 10.1159/000195745

Consortium, H. M. P. (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

De Goffau, M. C., Fuentes, S., Van Den Bogert, B., Honkanen, H., De Vos, W. M., Welling, G. W., et al. (2014). Aberrant gut microbiota composition at the onset of type 1 diabetes in young children. *Diabetologia* 57, 1569–1577. doi: 10.1007/s00125-014-3274-0

Fredricks, D. N. (2001). Microbial ecology of human skin in health and disease. *J. Investig. Dermatol. Symp. Proc.* 6, 167–169. doi: 10.1046/j.0022-202x.2001.00039.x

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi: 10.1126/science.1124234

Grenham, S., Clarke, G., Cryan, J. F., and Dinan, T. G. (2011). Brain-gut-microbe communication in health and disease. *Front. Physiol.* 2:94. doi: 10.3389/fphys.2011.00094

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Huang, W. K., Chang, J. W., See, L. C., Tu, H. T., Chen, J. S., Liaw, C. C., et al. (2012). Higher rate of colorectal cancer among patients with pyogenic liver abscess with *Klebsiella pneumoniae* than those without: an 11-year follow-up study. *Colorectal Dis.* 14, e794–e801. doi: 10.1111/j.1463-1318.2012.03174.x

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

Huang, Z. A., Chen, X., Zhu, Z., Liu, H., Yan, G. Y., You, Z. H., et al. (2017). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Islam, S. T., Srinivasan, S., and Craig, M. E. (2014). Environmental determinants of type 1 diabetes: a role for overweight and insulin resistance. *J. Paediatr. Child Health* 50, 874–879. doi: 10.1111/jpc.12616

Jung, J. W., Choi, J. C., Shin, J. W., Kim, J. Y., Park, I. W., Choi, B. W., et al. (2016). Lung microbiome analysis in steroid-nasmall yi, ukrainianve asthma patients by using whole sputum. *Tuberc. Respir. Dis.* 79, 165–178. doi: 10.4046/trd.2016.79.3.165

Laarhoven, T. V., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Larsen, N., Vogensen, F. K., Van Den Berg, F. W., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., et al. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* 5:e9085. doi: 10.1371/journal.pone.0009085

Li, M., Song, L. J., and Qin, X. Y. (2014). Advances in the cellular immunological pathogenesis of type 1 diabetes. *J. Cell Mol. Med.* 18, 749–758. doi: 10.1111/jcmm.12270

Liang, X., Li, H., Tian, G., and Li, S. (2014). Dynamic microbe and molecule networks in a mouse model of colitis-associated colorectal cancer. *Sci. Rep.* 4:4985. doi: 10.1038/srep04985

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe-disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

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.e3–352.e3. doi: 10.1016/j.jaci.2012.11.013

Murri, M., Leiva, I., Gomez-Zumaquero, J. M., Tinahones, F. J., Cardona, F., Soriguer, F., et al. (2013). Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study. *BMC Med.* 11:46. doi: 10.1186/1741-7015-11-46

Preston, J. A., Essilfie, A. T., Horvat, J. C., Wade, M. A., Beagley, K. W., Gibson, P. G., et al. (2007). Inhibition of allergic airways disease by immunomodulatory therapy with whole killed Streptococcus pneumoniae. *Vaccine* 25, 8154–8162. doi: 10.1016/j.vaccine.2007.09.034

Rayhan, F., Ahmed, S., Shatabda, S., Farid, D. M., Mousavian, Z., Dehzangi, A., et al. (2017). iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting. *Sci. Rep.* 7:17731. doi: 10.1038/s41598-017-18025-2

Sagar, S., Morgan, M. E., Chen, S., Vos, A. P., Garssen, J., Van Bergenhenegouwen, J., et al. (2014). *Bifidobacterium breve* and *Lactobacillus rhamnosus* treatment is as effective as budesonide at reducing inflammation in a murine model for chronic asthma. *Respir. Res.* 15:46. doi: 10.1186/1465-9921-15-46

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

Skov, L., and Baadsgaard, O. (2000). Bacterial superantigens and inflammatory skin diseases. *Clin. Exp. Dermatol.* 25, 57–61. doi: 10.1046/j.1365-2230.2000.00575.x

Sommer, F., and Backhed, F. (2013). The gut microbiota–masters of host development and physiology. *Nat. Rev. Microbiol.* 11, 227–238. doi: 10.1038/nrmicro2974

Soyucen, E., Gulcan, A., Aktuglu-Zeybek, A. C., Onal, H., Kiykim, E., and Aydin, A. (2014). Differences in the gut microbiota of healthy children and those with type 1 diabetes. *Pediatr. Int.* 56, 336–343. doi: 10.1111/ped.12243

Sun, K., Deng, H. J., Lei, S. T., Dong, J. Q., and Li, G. X. (2013). miRNA-338-3p suppresses cell growth of human colorectal carcinoma by targeting smoothened. *World J. Gastroenterol.* 19, 2197–2207. doi: 10.3748/wjg.v19.i14.2197

Sun, K., Wang, W., Zeng, J. J., Wu, C. T., Lei, S. T., and Li, G. X. (2011). MicroRNA-221 inhibits CDKN1C/p57 expression in human colorectal carcinoma. *Acta Pharmacol. Sin.* 32, 375–384. doi: 10.1038/aps.2010.206

Vael, C., Vanheirstraeten, L., Desager, K. N., and Goossens, H. (2011). Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol.* 11:68. doi: 10.1186/1471-2180-11-68

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Wu, C., Gao, R., Zhang, D., Han, S., and Zhang, Y. (2018). PRWHMDA: human microbe-disease association prediction by random walk on the heterogeneous network with PSO. *Int. J. Biol. Sci.* 14, 849–857. doi: 10.7150/ijbs.24539

Xue, Q., Sun, K., Deng, H. J., Lei, S. T., Dong, J. Q., and Li, G. X. (2014). MicroRNA-338-3p inhibits colorectal carcinoma cell invasion and migration by targeting smoothened. *Jpn. J. Clin. Oncol.* 44, 13–21. doi: 10.1093/jjco/hyt181

You, Z. H., Huang, Z. A., Zhu, Z., Yan, G. Y., Li, Z. W., Wen, Z., et al. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005455. doi: 10.1371/journal.pcbi.1005455

Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018). IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zhou, X. Z., Menche, J., Barabási, A., and Sharma, A. (2014). Human symptoms–disease network. *Nat. Commun.* 5:4212. doi: 10.1038/ncomms5212

Conflict of Interest Statement: 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.

Copyright © 2018 Peng, Yin, Zhou, Liu and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 23 October 2018
doi: 10.3389/fmicb.2018.02500

Predicting Influenza Antigenicity by Matrix Completion With Antigen and Antiserum Similarity

Peng Wang[1], *Wen Zhu*[2], *Bo Liao*[1,2]*, *Lijun Cai*[1]*, *Lihong Peng*[3] and *Jialiang Yang*[2,4]*

[1] College of Information Science and Engineering, Hunan University, Changsha, Changsha, China, [2] School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [3] School of Computer Science, Hunan University of Technology, Zhuzhou, China, [4] Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine At Mount Sinai, New York, NY, United States

The rapid mutation of influenza viruses especially on the two surface proteins hemagglutinin (HA) and neuraminidase (NA) has made them capable to escape from population immunity, which has become a key challenge for influenza vaccine design. Thus, it is crucial to predict influenza antigenic evolution and identify new antigenic variants in a timely manner. However, traditional experimental methods like hemagglutination inhibition (HI) assay to select vaccine strains are time and labor-intensive, while popular computational methods are less sensitive, which presents the need for more accurate algorithms. In this study, we have proposed a novel low-rank matrix completion model MCAAS to infer antigenic distances between antigens and antisera based on partially revealed antigenic distances, virus similarity based on HA protein sequences, and vaccine similarity based on vaccine strains. The model exploits the correlations of viruses and vaccines in serological tests as well as the ability of HAs from viruses and vaccine strains in inferring influenza antigenicity. We also compared the effects of comprehensive 65 amino acids substitution matrices in predicting influenza antigenicity. As a result, we applied MCAAS into H3N2 seasonal influenza virus data. Our model achieved a 10-fold cross validation root-mean-squared error (RMSE) of 0.5982, significantly outperformed existing computational methods like antigenic cartography, AntigenMap and BMCSI. We also constructed the antigenic map and studied the association between genetic and antigenic evolution of H3N2 influenza viruses. Finally, our analyses showed that homologous structure derived amino acid substitution matrix (HSDM) is most powerful in predicting influenza antigenicity, which is consistent with previous studies.

Keywords: hemagglutination inhibition assay, low-rank matrix completion, influenza antigenicity, antigenic map, HA protein sequence information

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China
Hauke Busch,
Universität zu Lübeck, Germany

***Correspondence:**
Jialiang Yang
jialiang.yang@mssm.edu
Bo Liao
dragonbw@163.com
Lijun Cai
ljcai@hnu.edu.cn

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

Received: *30 May 2018*
Accepted: *01 October 2018*
Published: *23 October 2018*

Citation:
Wang P, Zhu W, Liao B, Cai L, Peng L and Yang J (2018) Predicting Influenza Antigenicity by Matrix Completion With Antigen and Antiserum Similarity. Front. Microbiol. 9:2500. doi: 10.3389/fmicb.2018.02500

INTRODUCTION

According to the United States Centers for Disease Control and Prevention (CDC), seasonal influenza and its linked respiratory diseases cause approximately 650,000 deaths annually worldwide, posing a serious threat to human health and socio-economic environment (WHO, 2017). This is mainly attributed to seasonal influenza viruses that frequently evade immunity in

the human population through mutations in their hemagglutinin (HA) and neuraminidase (NA) surface glycoproteins (Hay et al., 2001; Neher et al., 2016). The most effective way to prevent influenza virus infection is to inoculate vaccines with similar antigenicity to the influenza virus (Sun et al., 2013). Therefore, timely and accurate identification of the effectiveness of existing vaccines on circulating virus strains is critical for vaccine design and influenza surveillance (Smith et al., 2004; Huang et al., 2017). However, the task is challenging (Blackburne et al., 2008; Yao et al., 2017). To facilitate the selection and design of vaccine strains, the World Health Organization's (WHO) Global Influenza Surveillance and Response System (GISRS) continuously monitors genotypic and antigenic characteristics of circulating viruses (Barr Ig, 2014).

The hemagglutination inhibition (HI) is one of the most popular experimental methods for measuring the effectiveness of a vaccine against an influenza virus (Hirst, 1943). It is a binding assay used to characterize the ability of antisera (vaccines) to block HA of antigens (viruses) from agglutinating red blood cells (RBC). Based on the HI assay, the concept of antigenic distance can be used to quantitatively describe the closeness among antigens. The antigenic distance is often defined to be the Euclidean distance between their representing vectors in a normalized HI table according to multiple reference antisera (Cai et al., 2010). HI assay and its derived antigenic distance provide great convenience for comparing antigenic similarity among influenza viruses (Fouchier et al., 2010; Neher et al., 2016). However, HI assays are expensive and time-consuming, so it is impractical to use it to measure the antigenic similarity among all antigens and antisera (Sun et al., 2013). This urges the need to explore effective computational methods (Liao et al., 2012a,b; Chen et al., 2018) to estimate the antigenic distance between an antigen and an antiserum (Liao et al., 2010, 2015a,b; Li et al., 2013; Liang et al., 2016; Peng et al., 2017).

A popular category of methods for predicting the antigenicity of influenza virus is the sequence-based method. Unlike imputation-based methods, sequence-based methods often explore the association between mutations in the HA protein and antigenic distances obtained from serological tests (Lee and Chen, 2004; Barnett et al., 2012; Li et al., 2016). The antigenic difference between two influenza viruses indicates whether they antigenic variant, which is measured by either an antigenic distance or simply a binary value (Lee and Chen, 2004; Smith et al., 2004; Liao et al., 2008). For example, a model based on multiple logistic regression was proposed by Liao et al. to predict antigen variants. For further exploration, 65 amino acid substitution models based on 20 amino acid physicochemical groups were also studied. The experimental results showed that high agreement was achieved in the H3N2 influenza data from 1999 to 2003 (Liao et al., 2008). Huang et al. introduced a decision tree algorithm to predict antigenic variants (Huang et al., 2009). Sun et al. proposed a bootstrapped ridge regression model consisting of antigenic related sites, which uses the quantitative amino acid substitutions in the HA1 [a sub-unit of HA forming globular domain (Wang et al., 2015)] protein sequence to predict antigenic distances (Sun et al., 2013). Inspired by the co-evolution of HA1 that may have contributed to antigen evolution, Yang et al. integrated the single mutation and co-mutation characteristics of the HA1 sequence and proposed a Lasso model (Yang et al., 2014). Neher et al. proposed an optimization model for interpreting known antigen data and studied its ability to predict future influenza virus population composition (Neher et al., 2016). However, these methods rely on the reliability of rapidly changing antigen-associate sites (Sun et al., 2013).

Imputation-based methods are widely used for predicting and visualizing the antigenicity of influenza viruses (Smith et al., 2004; Cai et al., 2010; Barnett et al., 2012). They are based on the assumption that the antigens and antisera are located in a low dimensional space (i.e., the normalized HI table is of low rank), so the HI table can be fully recovered from partially revealed HI titers (Lapedes and Farber, 2001). For example, Smith et al. proposed antigenic cartography for visualizing and predicting antigenic evolution of influenza viruses (Smith et al., 2004). They first transformed the known values in the HI table to Euclidean distances and then embedded them into a 2D map using the modified multidimensional scaling (MDS) method. This antigenic map implicitly implies the distance between antigen and antiserum with unknown HI titer. Cai et al. first recovered the normalized HI table by a low-rank matrix completion method (Cai et al., 2010), and then calculated the antigenic distance using the fully recovery normalized HI table and mapped it into a 2D or 3D (Barnett et al., 2012) antigenic map. Imputation-based methods can better detect the antigenic evolutionary trend of H3N2 influenza virus, but it is still insufficient. For example, the accuracy of its predicted antigenic distance is yet to be improved (Huang et al., 2017).

The antigenic evolution of influenza viruses are ultimately caused by genetic changes of the viruses especially on HA and NA genes, thus principally the sequence information of antigens and antisera will help predict missing values in HI. In this study, we propose a novel algorithm called matrix completion with antigen and antiserum similarity (MCAAS), which integrates antigen sequence information and antiserum information in a low-rank matrix completion model to predict influenza antigenicity. To our best knowledge, this the first model to leverage both the low-rank space of viruses spaces and the importance of genetic mutations in predicting influenza antigenicity. To explore the influence of different amino acids properties on the prediction of the antigenicity of the H3N2 influenza virus, we systematically compared the 65 amino acid substitution matrices in the AAindex database (Shuichi Kawashima et al., 2008), reflecting a comprehensive list of amino acid properties, including structural, physicochemical, and biochemical information. In addition, in order to make full use of the information, we have proposed a mixed-rank strategy to improve the sliding window method. The algorithm proposed in this paper was applied to H3N2 influenza data from 1968 to 2003. We then constructed an antigenic map based on the fully recovered HI table and evaluated existing vaccine strains. Finally, we explored the relationship between the genetic and antigenic evolution of the influenza virus in H3N2 data.

MATERIALS AND METHODS

Dataset and Problem Formulation

H3N2 influenza data are used in this study (Smith et al., 2004), which is a partially revealed HI table consisting of 253 viruses (antigens) and 79 vaccine (antisera) from 1967 to 2003, i.e., a matrix of 253 rows and 79 columns. The HI table contains Type I data, Type II data, Type III data, which are regular HI titers, low reactors (i.e., the HI titers less than a threshold) and missing values (Cai et al., 2010). Similar to many literatures (Smith et al., 2004; Cai et al., 2010; Sun et al., 2013; Huang et al., 2017), the HI table was normalized to facilitate subsequent analyses. We also downloaded HA protein sequences of viruses and vaccine strains related to HI tables from the NCBI influenza database. Only the sequence on 329 sites belonging to the HA1 protein was kept for further analysis (Yao et al., 2017). We also downloaded 65 amino acid substitution matrices from the AAindex database (Shuichi Kawashima et al., 2008) to analyze the effect of amino acid structure, physical and biological information on predicting influenza antigenicity. In this paper, the problem is how to accurately estimate low reactors and predict missing values based on values on regular entries and fusion information, which combines multiple amino acids substitution matrices and sequence information of the viruses and vaccine strains.

Matrix Completion With Antigen and Antiserum Similarity

In this paper, we consider the problem of predicting the antigenicity of influenza viruses against vaccines, which is to fill the missing values in the HI table (as well as corrections for Type I and Type II data). Without considering the temporal bias effect, we can convert this problem into a matrix completion problem (Cai et al., 2010). Specifically, we use H to denote an HI table with m rows and n columns, which corresponds to m antigens and n antisera. Let E to represent the corresponding Type I and Type II data locations in H. Let X be the underlying matrix to recover H. Since X is in a low-dimensional space for influenza viruses, we assume that $r \ll \min(m, n)$ as the rank of X. For some $\Sigma_{r\times r}$ matrices, X can be expressed as $X = U_{m\times r}\Sigma_{r\times r}(V_{n\times r})^T$ according to singular value decomposition.

In the literature (Huang et al., 2017), it has been shown that incorporating the HA protein sequence information of viruses into the matrix completion method achieves better results. However, since the model does not use vaccine strains HA protein sequence information, the use of information is incomplete. Moreover, the calculation of protein sequence similarity in this model does not take into account the physicochemical and biochemical properties of amino acids. In addition, the effect of the antigenic determinant regions on protein properties was not discussed in detail. Therefore, in order to solve the above deficiencies, we propose two new models that incorporate the above information into the matrix completion model.

Model 1 without Type II data

$$\mathrm{Min}_X \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}^E - H_{ij}^E\right)^2 + \lambda_1 G(X) + \lambda_2 \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} K_{ij}\left\|X^i - X^j\right\|^2 + \lambda_3 \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} T_{ij}\left\|\left(X^T\right)^i - \left(X^T\right)^j\right\|^2$$

Model 2 with Type II data

$$\mathrm{Min}_X \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}^E - H_{ij}^E\right)^2 I(X_{ij}^E \geq \theta_{ij}) + \lambda_1 G(X) + \lambda_2 \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} K_{ij}\left\|X^i - X^j\right\|^2 + \lambda_3 \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} T_{ij}\left\|\left(X^T\right)^i - \left(X^T\right)^j\right\|^2$$

The function $G(X) = \sum_{i=1}^{m} g\left(\frac{\|U^i\|^2}{3\delta r}\right) + \sum_{i=1}^{n} g\left(\frac{\|V^i\|^2}{3\delta r}\right)$ is a regularization term, where $g(z) = e^{(z-1)^2} - 1$ when $z \geq 1$ and $g(z) = 0$, otherwise. $U^i$and V^i denote the *ith* row of U and V, respectively and $\delta = \max(m, n)$ (Keshavan et al., 2009a; Cai et al., 2010). K_{ij} is the HA protein sequence similarity between virus i and j, T_{ij} is the HA protein sequence similarity between vaccine strains for vaccine i and j. X^i and X^j represent the *ith* row and *jth* row of X, respectively. $(X^T)^i$ and $(X^T)^j$ represent the *ith* column and *jth* column of X, respectively. The three parameters λ_1, λ_2, and λ_3 control the contribution of matrix completion, HA1 protein sequence of antigens and HA1 protein sequence of vaccine strains to recover the matrix. The third and fourth terms in the model are based on the assumption that if the viruses (vaccine strains) have similar HA protein sequences (especially in antigenic determinant regions), they should have similar HI titers against the same group of vaccines (viruses). Based on previous literatures, the antigenic regions B and C seems to be more important than A, D, and E (Yao et al., 2017). Thus, we define $K_{ij} = \xi_1 K_{ij}^{ADE} + \xi_2 K_{ij}^{BC} + K_{ij}^{other}$($T_{ij} = \xi_1 T_{ij}^{ADE} + \xi_2 T_{ij}^{BC} + T_{ij}^{other}$) as the similarity calculation formula, in which ξ_1 and ξ_2 are the parameter to control the weight of antigenic determinants. K_{ij}^{ADE} measures sequence similarity on antigenic determinant regions site A, site D, and site E. K_{ij}^{BC} measures sequence similarity on antigenic determinant regions site B and site C. K_{ij}^{other} measures sequence similarity on other site. Parameters λ_1, λ_2, λ_3, ξ_1, and ξ_2 were tuned by 10-fold cross-validation.

An Alternating Gradient Descend Method

To solve Model 1, we propose an alternating gradient descend AGD method similar to literature (Keshavan et al., 2009a; Cai et al., 2010). Since the corresponding singular vectors are highly concentrated on the high-weight row (column) index when $|E| = \Theta(n)$ (Keshavan et al., 2009b), in order to ensure that the number of non-zero values per row (column) is less than $\frac{2|E|}{m}$($\frac{2|E|}{n}$), we need to trim the H matrix. When a row (column) has more non-zero values than $\frac{2|E|}{m}$($\frac{2|E|}{n}$), we randomly set some non-zero values to zero.

We replace all missing values in H with 0 to form $H^{(0)}$. After singular value decomposition (SVD) $H^{(0)} = U\Sigma V^T$, we set $U^{(0)} = U_0 * \sqrt{m}$ and $V^{(0)} = V_0 * \sqrt{n}$ as initial values, where, U_0 and V_0 consist of the first r columns of U and V, respectively.

Then we use the following updates until convergence or reaching a preset maximum number of iterations.

Fix $U^{(t)}$ and $V^{(t)}$ and calculate the matrix $\Sigma_{r\times r}$ to minimize the Model 1 as follows:

$vec(\Sigma_{r\times r}) = \left(V^TV \otimes V^TH^TV + \lambda_2 V^TV \otimes U^TK_L^TU + \lambda_3 U^TU \otimes V^TT_L^TV\right)^{-1} vec(U^THV)$

where $\otimes$ is Kronecker Product, K_L is the Laplacian matrix of K and T_L is the Laplacian matrix of T.

Update $U^{(t+1)}$ and $V^{(t+1)}$ using gradient descent: $U^{(t+1)} = U^{(t)} + \alpha\nabla U^{(t)}$ and $V^{(t+1)} = V^{(t)} + \alpha\nabla V^{(t)}$.

The gradients of U and V are:

$$\nabla U = \left(\left(U\Sigma V^T\right) - H\right)V\Sigma^T + UQ_U + \lambda_1 f\left(U, 2e^{(Q_{U1}-I_1)^2}(Q_{U1}-I_1)\right) + 4\lambda_2 AU\Sigma\left(V^TV\right)\Sigma^T + 4\lambda_3(U\Sigma V^T)BV\Sigma^T$$

$$\nabla V = \left(\left(U\Sigma V^T\right) - H\right)^T U\Sigma + VQ_V + \lambda_1 f\left(V, 2e^{(Q_{V1}-I_2)^2}(Q_{V1}-I_2)\right) + 4\lambda_2 V\Sigma^TU^TAU\Sigma + 4\lambda_3 BV\Sigma^T(U^TU)\Sigma$$

Where $A = \left[a_{ij}\right]_{m\times m}$ and $B = \left[b_{ij}\right]_{n\times n}$ are symmetric matrix with $a_{ij} = \begin{cases}\Sigma_{h\neq i}K_{ih} & if\ i=j\\ -K_{ij} & if\ i\neq j\end{cases}$ and $b_{ij} = \begin{cases}\Sigma_{h\neq i}T_{ih} & if\ i=j\\ -T_{ij} & if\ i\neq j\end{cases}$

$$Q_U = \frac{1}{m}U^T\left(\left(H - \left(U\Sigma V^T\right)\right)\right)V\Sigma^T,$$

$$Q_V = \frac{1}{n}V^T\left(W \circ \left(H - \left(U\Sigma V^T\right)\right)\right)^T U\Sigma$$

$$I_1 = (1,1,\ldots 1)^T_{m\times 1}, I_2 = (1,1,\ldots 1)^T_{n\times 1},$$

$$Q_{U1} = \frac{1}{3\alpha r}\begin{bmatrix}\sum_{j=1}^{r}U_{1j}^2\\ \sum_{j=1}^{r}U_{2j}^2\\ \vdots\\ \sum_{j=1}^{r}U_{mj}^2\end{bmatrix}; Q_{V1} = \frac{1}{3\alpha r}\begin{bmatrix}\sum_{j=1}^{r}V_{1j}^2\\ \sum_{j=1}^{r}V_{2j}^2\\ \vdots\\ \sum_{j=1}^{r}V_{nj}^2\end{bmatrix}$$

Where $\alpha = \max(m,n)$ and $f(C_{m\times r}, D_{m\times 1}) = Z_{m\times r}$ with

$$Z_{ij} = \begin{cases}\frac{1}{\alpha r}C_{ij} * D_i & if\ D_i > 0\\ 0 & otherwise\end{cases}$$

The difference between Model 2 and Model 1 is that Model 2 has Type II data, where the Type II data is treated differently by multiplying B in the model. Therefore, we use the same method to solve Model 1 and Model 2. We only need to replace $\sum\sum\left(X_{ij}^E - H_{ij}^E\right)^2$ with $\sum\sum\left(X_{ij}^E - H_{ij}^E\right)^2 I(X_{ij}^E \geq \theta_{ij})$ and replace $\left(\left(U\Sigma V^T\right)^E - H^E\right)$ with $\left(\left(U\Sigma V^T\right)^E - H^E\right)\cdot$ I. Here, I is an indication matrix, dot multiplication denotes the multiplication of corresponding elements between the matrices.

A Sliding Window Method

Since there is temporal bias in the HI matrix that can affect the accuracy of the matrix completion, in this paper we introduce a sliding window method to reduce this effect. The method mainly based on the principle that the temporal bias effect becomes smaller in the temporal-grouped submatrix than in the entire HI matrix. The generally flow of the method is summarized in **Figure 1**: let Y_0 and Y be the starting and ending year and W be the window size. Then the $i+1th$ window year span should be from $(Y_0 + i)$ to $(Y_0 + i + W)$ and there is a total of $(Y - W + 1)$ windows. Since the rank of the submatrix is less than or equal to the rank of the full-matrix, it is reasonable to consider a mixed rank rather than a single rank consistent with the full-matrix. In this paper, the rank of the submatrix is set to $rank'$ and $(rank' - 1)$ in the window sliding method, and $rank'$ is the setting of the full-matrix rank. Missing values are estimated on each submatrix in the case of rank $rank'$ and $(rank' - 1)$, and then the average of these estimates is taken as the recovered value of the corresponding position of the matrix. After the window is sliding, a partially recovered HI matrix is obtained, and on the basis of this, the algorithm proposed in this paper is performed on the whole window to fill in the values that has not been recovered.

Performance Evaluation

The performance of imputation algorithms is evaluated using the root-mean-squared error (RMSE). Given k values $\{O_1, O_2, \ldots, O_k\}$ and $\{P_1, P_2, \ldots, P_k\}$, the RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{k}(O_i - P_i)^2}{k}} \quad (1)$$

Where O_i represents an observed value and P_i represents the corresponding predicted value. The smaller the RMSE value is, the closer the predicted value is to the observed value, indicating that the performance of the algorithm is better.

In this paper, we use 10-fold cross validation to calculate the RMSE value. Specifically, the H matrix is randomly divided into 10 equal parts. We will run it repeatedly for 10 times in the experiment until each part was used as the prediction set once. Each time, we use 9 parts for matrix completion; then calculate the RMSE between the completed matrix and the observed matrix entry in the remaining part. The mean RMSE between the predicted values and observed values across 10 runs are used to compare different methods. And the model parameters $\lambda_1, \lambda_2, \lambda_3$, ξ_1, ξ_2, r, and w are tuned in this process.

Construction of Antigenic and Genetic Cartography

Similar to literature (Cai et al., 2010; Barnett et al., 2012), we use the Euclidean distance between viruses after completion of the matrix as the antigenic distance. Then, multidimensional scaling (MDS) is used to generate virus coordinates and construct antigen maps based on antigenic distances. The construction of the genetic map is similar to the antigenic map. We first calculate the P-distance matrix between pairs of viruses and use MDS to construct the genetic map.

FIGURE 1 | A cartoon to show the sliding window process. Row and column indicate antigen and antiserum, respectively, which are placed chronologically from the up-left to bottom-right of the window. MCAAS is used from the first window consisting of antigens and antisera starting from Year1 to the t-w+1 window consisting of antigens and antisera starting from Year t-w+1. In the shaded region, the final values are taken as the mean of the completed values in all related windows. In the end, MCAAS is performed on the whole window.

RESULTS

Dataset

In this paper, we used the H3N2 influenza data as our test dataset for HI values from 253 viruses against 79 antisera. There are 3,991 observed HI values in this matrix, and the sparseness is about 0.2. These viruses formed 11 antigenic clusters, namely HK68, EN72, VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97, and FU02 (Smith et al., 2004). The HA1 protein sequences of viruses and vaccine strains were then downloaded from the NCBI Influenza Virus Database.

Matrix Completion for HI Table of H3N2

In this paper, we used the similarity matrix between protein sequences to assist matrix completion. There are many amino acid substitution matrices that reflect different amino acid properties, and the literatures (Lee and Chen, 2004; Liao et al., 2008) show that the substitution matrix is critical to the accuracy of the prediction. To investigate the effect of different amino acid properties on the evolution of antigens, we used the method in this article to evaluate 65 amino acid substitution matrices with parameters set to $\xi_1 = 500$, $\xi_2 = 1000$, $r = 10$, $w = 32$, lam1 = 1E-4, and lam2 = lam3 = 2.5E-7 (after normalizing the similarity matrix). The 10-fold cross-validation root mean square errors (RMSE) for all 65 substitution matrices were presented in **Supplementary Table S1**, with the top 12 RMSEs summarized in **Table 1**.

As can be seen from **Table 1**, different substitution matrices have a certain influence on the prediction result. The best substitution matrix is "Homologous structure derived matrix (HSDM) for alignment of distantly related sequences." The RMSE obtained by using it is 0.6349. This implies the importance of HA1 protein structure in influenza antigenicity, since the best ones are based on structure-based substitution matrices, which is very reasonable because the structural information is the key to

TABLE 1 | The top 12 amino acids substitution matrices in predicting influenza antigenicity.

Accession number	Description	RMSE
PRLA000102	Homologous structure derived matrix (HSDM) for alignment of distantly related sequences	0.6349
HENS920101	BLOSUM45 substitution matrix	0.6351
JOND920103	The 250 PAM PET91 matrix	0.6352
QU_C930101	Cross-correlation coefficients of preference factors main chain	0.6352
PRLA000101	Structure derived matrix (SDM) for alignment of distantly related sequences	0.6352
KANM000101	Substitution matrix (OPTIMA) derived by maximizing discrimination between homologs and non-homologs	0.6352
CSEM940101	Residue replace ability matrix	0.6353
LUTR910107	Structure-based comparison table for other class	0.6354
MIYS930101	Base-substitution-protein-stability matrix	0.6354
BENS940104	Genetic code matrix	0.6355
NIEK910102	Structure-derived correlation matrix 2	0.6355
HENS920103	BLOSUM80 substitution matrix	0.6358

TABLE 2 | Ten-fold cross-validation RMSEs for different window sizes and ranks on H3N2 data.

w\r	6	7	8	9	10	11	12	13	14
8	1.2563	1.3737	1.5396	/	/	/	/	/	/
12	1.0394	1.0995	1.1398	1.1838	1.1749	1.1676	1.4021	/	/
16	0.9902	0.9802	0.9666	0.9011	0.9427	0.9909	1.1016	1.1051	1.2234
20	0.9434	0.9125	0.8681	0.9542	0.8833	0.8621	0.8765	0.9284	0.9421
24	0.8322	0.8705	0.8218	0.8648	0.7294	0.7396	0.7475	0.7722	0.8516
28	0.7939	0.6995	0.7879	0.7894	0.6792	0.6320	0.6881	0.7083	0.7958
32	0.7856	0.7294	0.6503	0.5982	0.6068	0.6872	0.7069	0.6779	0.6577

TABLE 3 | Ten-fold cross-validation RMSEs for the analysis of single virus information dependence on H3N2 data.

row	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10	0.6028	0.5985	0.6052	0.5966	0.6288	0.6385	0.6427	0.6410	0.6268	0.6585
130	0.6049	0.6360	0.6182	0.6583	0.6453	0.6693	0.6746	0.6171	0.6157	0.7115
240	0.5803	0.5762	0.5862	0.5886	0.5933	0.6115	0.6176	0.6405	0.6438	0.7651

"row" means the row where the information was deleted in HI. "10%, ... , 100%" means the proportion of information deleted.

TABLE 4 | Ten-fold cross-validation RMSEs for the analysis of combined virus information dependence on H3N2 data.

row	RMSEs	row	RMSEs	row	RMSEs	row	RMSEs
10	0.6585	130	0.7115	10/30	0.7296	10/30/50/70	0.9707
30	0.6987	150	0.7874	50/240	0.8178	90/110/130/150	0.9405
50	0.6958	180	0.7282	70/110	0.8524	180/200/220/240	1.0118
70	0.7742	200	0.7666	90/220	0.7988	10/30/50/70/90/110	1.0101
90	0.7131	220	0.7280	130/180	0.7419	130/150/180/200/220/240	1.1963
110	0.7543	240	0.7651	150/200	0.8243	10/30/50/70/90/110 130/150/180/200/220/240	1.4959

"row" means the row where the information is completely deleted in HI. For example, "10" and "30" mean that the 10th line and the 30th line are deleted, respectively. "10/30" means the 10th line and the 30th line are deleted at the same time.

the binding affinity between the antigen and the antisera (Hirst, 1943).

We set the mixed low-rank r to vary from 6 to 14, and the sliding window size W to vary from 8 to 32 with a step size of 4. Here we used the PRLA000102 substitution matrix to measure sequence similarity and use the mixed rank window sliding method proposed in this paper. Other parameters are consistent with before. We listed the 10-fold cross validation RMSEs for different r and W in **Table 2**. As can be seen from the table, the lowest RMSE 0.5982 is achieved at window size 32 and rank 9 (mixed rank of 8 and 9). According to Huang et al. (2017), the best RMSE value for BMCSI is 0.6586 and those for antigenic cartography (Smith et al., 2004; Cai et al., 2010) and AntigenMap (Barnett et al., 2012) are 1.04 and 1.05, respectively. The above results indicate that the complete HA1 protein sequence information with discriminating antigenic determinant regions is a good compensation for low rank matrix completion. From **Tables 1**, **2**, it is clear that a mixed rank sliding window method is more appropriate in completing the H3N2 influenza data.

In order to analyze the dependence of the model on available virus information, we selected 3 viruses for single virus information analysis, and selected 12 viruses for combined virus information analysis. In the analysis of single virus information dependence, we selected the 10th, 130th, and 240th rows of virus information in the HI table, sequentially deleted about 10% of the information. The results of the analysis were shown in **Table 3**. It can be seen from **Table 3** that with the increase of information deletion, the prediction performance of the model is generally declining. In the analysis of combined virus information dependence, we ran all the cases where all the individual virus information was deleted, the case of 2 virus combinations, the case of 4 virus combinations, the case of 6 virus combinations, and the case of 12 virus combinations. The results of the analysis were shown in **Table 4**. As can be seen from **Table 4**, as more virus information is deleted, the prediction effect becomes worse and worse.

Antigenic Cartography for H3N2 Viruses

Based on the antigenic distance predicted by the MCAAS method, we constructed an antigenic map of 253 viruses in H3N2 by multidimensional scaling in **Figure 2**. As can be seen from **Figure 2**, 11 antigen clusters can be distinguished very well, especially VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97, and FU02. It is reasonable since there are more HI observations in these later years, resulting in more reliable calculations. We can also find from **Figure 2** that the virus has generally evolved locally along an S-shaped pathway, which are consistent with previous research (Smith et al., 2004).

We also calculated the average antigenic distances within cluster and between clusters (see **Table 5**), which are generally consistent with **Figure 2**. For example, the antigenic distance between BE92 and BE89 is much greater than the antigenic distance between SI87 and BE89 in **Figure 2**, whose

corresponding average antigenic distances were 4.86 and 3.23, respectively. From **Table 3**, we can find that the average within-cluster distances of the 11 clusters are all <1.7, and the inter-cluster distances are >1.7 except for BK79-BK79 (1.64) and BE92-BE92 (1.68). In addition, the antigenic distance between viruses becomes larger as the time interval increases.

Relationship Between Influenza Genetic and Antigenic Evolution of H3N2

To further explore relationship between the genetic and antigenic evolution of the H3N2 virus, we not only constructed a genetic map (**Figure 3**) of the 253 viruses using the uncorrected P-distance and MDS, but also calculated the average genetic distance (uncorrected P distance) of viruses within and between 11 antigen clusters (**Table 6**). As shown in **Figure 3**, it can be seen that the genetic evolution of the virus proceeds along a semicircle. By comparing **Figure 2** with **Figure 3**, we found that the genetic and antigenic profiles are partially consistent. However, their evolutionary shapes are different, and genetic maps are more continuous, while antigenic maps are more punctual. From **Table 4**, we can see that the genetic distance between clusters increases with the increase of time span. The average genetic distance within-cluster varies from 0.004 to 0.025, while the average genetic distance between-clusters varies from 0.025 to 0.165.

Although the genetic map and the antigenic map are roughly consistent, we also found that some viruses are very close in the genetic map, but are far apart in the antigenic map. For example, BE89 and BE92 are very close in the genetic map (**Figure 3**) with the average genetic distance only 0.043, but they are far in the antigenic map (**Figure 2**) with the average antigenic distance 4.86. This shows that not all genetic changes are equivalent to cause antigenic changes and different protein sites contribute

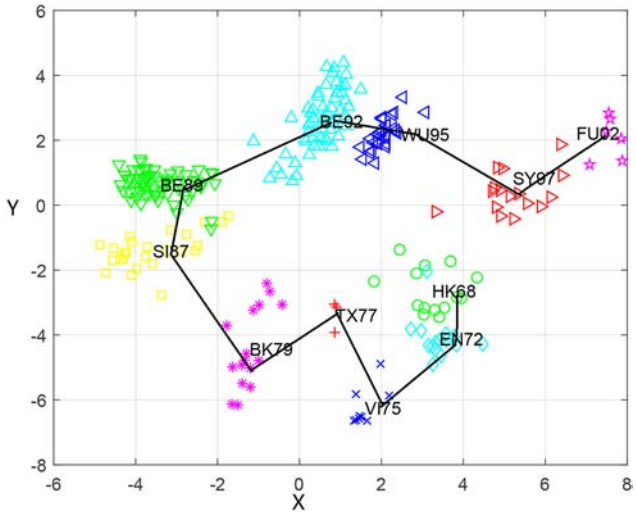


FIGURE 2 | The antigenic cartography of H3N2 influenza viruses from 1968 to 2003 constructed by MCAAS. Each node denotes a virus and the distance between two nodes reflect their antigenic distance. The viruses in 11 antigenic clusters (HK68, EN72, VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97, and FU02) are marked with different shapes and colors.

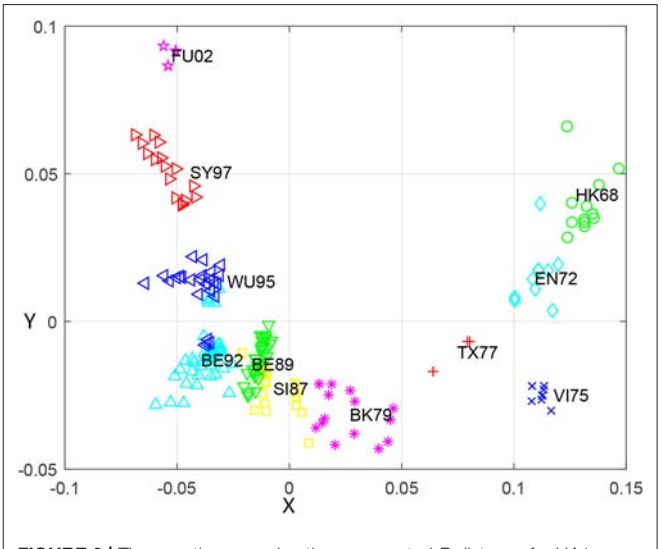


FIGURE 3 | The genetic map using the uncorrected-P distance for HA1 protein sequences of H3N2 influenza virus from 1968 to 2003. Each node denotes a virus and the distance between two nodes reflect their genetic distance. The viruses in 11 antigenic clusters (HK68, EN72, VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97, and FU02) are marked with different shapes and colors.

TABLE 5 | The average antigenic distances among viruses within and between 11 antigenic clusters for H3N2 influenza.

	HK68	EN72	VI75	TX77	BK79	SI87	BE89	BE92	WU95	SY97	FU02
HK68	1.29	1.97	5.00	4.41	6.48	7.27	6.83	4.80	6.15	5.55	7.00
EN72	1.97	0.68	2.12	3.18	6.45	7.68	7.57	6.52	7.01	5.04	8.18
VI75	5.00	2.12	0.52	1.94	3.71	8.49	7.07	10.48	7.77	6.57	16.95
TX77	4.41	3.18	1.94	0.28	1.77	5.08	6.35	4.57	4.32	6.07	11.65
BK79	6.48	6.45	3.71	1.77	1.64	4.44	6.00	6.62	4.51	8.70	16.96
SI87	7.27	7.68	8.49	5.08	4.44	1.40	3.23	4.61	7.14	7.98	12.39
BE89	6.83	7.57	7.07	6.35	6.00	3.23	0.88	4.86	5.21	7.01	9.96
BE92	4.80	6.52	10.48	4.57	6.62	4.61	4.86	1.68	2.55	6.19	8.03
WU95	6.15	7.01	7.77	4.32	4.51	7.14	5.21	2.55	0.85	3.76	5.79
SY97	5.55	5.04	6.57	6.07	8.70	7.98	7.01	6.19	3.76	1.22	2.32
FU02	7.00	8.18	16.95	11.65	16.96	12.39	9.96	8.03	5.79	2.32	1.13

TABLE 6 | The average genetic distances among viruses within and between 11 antigenic clusters for H3N2 influenza.

	HK68	EN72	VI75	TX77	BK79	SI87	BE89	BE92	WU95	SY97	FU02
HK68	0.022	0.04	0.077	0.074	0.113	0.13	0.132	0.143	0.138	0.157	0.165
EN72	0.043	0.02	0.051	0.047	0.089	0.11	0.116	0.126	0.125	0.14	0.15
VI75	0.077	0.05	0.007	0.044	0.084	0.109	0.118	0.126	0.13	0.144	0.154
TX77	0.074	0.05	0.044	0.011	0.056	0.08	0.094	0.098	0.105	0.127	0.138
BK79	0.113	0.09	0.084	0.056	0.024	0.042	0.057	0.07	0.085	0.109	0.12
SI87	0.13	0.11	0.109	0.08	0.042	0.015	0.025	0.042	0.059	0.085	0.109
BE89	0.132	0.12	0.118	0.094	0.057	0.025	0.012	0.043	0.05	0.074	0.101
BE92	0.143	0.13	0.126	0.098	0.07	0.042	0.043	0.021	0.037	0.072	0.099
WU95	0.138	0.13	0.13	0.105	0.085	0.059	0.05	0.037	0.022	0.057	0.085
SY97	0.157	0.14	0.144	0.127	0.109	0.085	0.074	0.072	0.057	0.025	0.051
FU02	0.165	0.15	0.154	0.138	0.12	0.109	0.101	0.099	0.085	0.051	0.004

differently to antigenic evolution (Smith et al., 2004; Lee et al., 2016).

DISCUSSIONS

It is known that the antigenicity of influenza virus changes very quickly. To prevent influenza outbreaks caused by changes in influenza virus antigens, the 80 WHO collaborating laboratories actively monitored the influenza viruses to determine vaccine strains for the next flu season. However, the selection of influenza vaccine strains is a labor-intensive and time-consuming process that relies on the identification of antigenic variants. In this paper, we propose a new method for integrating similarity information between viruses and between vaccines into matrix completion. The completed matrix was also used for constructing antigenic map, which helps to select vaccine strains.

With the development of sequencing technology, the acquisition of sequence information becomes easier. In the literature (Huang et al., 2017), it is shown that the integration of sequence information contributes to the prediction of viral antigenicity. This paper further explores the effect of fusion of sequence information on the prediction of virus antigenicity, mainly from four perspectives. (1) The integration of sequence information improved antigenic prediction. Not only the similarity information of the virus sequences but also the similarity information of the vaccine strains was used. (2) We discussed in more detail the influence of antigenic determinant regions on antigenic changes and further analyzed the B and C regions in the five antigenic determinant regions. (3) We analyzed 65 substitution matrices, which reflect the different physicochemical and biochemical properties of amino acids. The results show that the characteristics of the structure have a greater impact on antigen evolution. (4) We proposed a mixed rank sliding window method that can solve matrix completion problems more reasonably than single rank methods. As a result, our method reduces the prediction RMSE compared with the literature (Huang et al., 2017) and previous interpolation methods (Smith et al., 2004; Cai et al., 2010). On this basis, we also discovered a semi-circular genetic evolution and S-shaped antigen evolution, which is consistent with previous findings (Smith et al., 2004; Fouchier et al., 2010).

It is worth noting that although we used the H3N2 data in this paper, our method is applicable to all influenza subtype data such as H1N1, H5N1, and H7N9. In fact, this method could be applied to any data with a response matrix and predictive characteristics, such as the prediction of diseases and drugs, the association between miRNAs with diseases, and the recognition of protein folds (Wei and Zou, 2016). For example in drug-response prediction, the entries in the matrix represent the effect of drugs on samples, which can be formulated as a typical matrix completion problem. We believe that the similarity among drugs based on their chemical properties and the samples genetic and gene expression similarity will also help to infer drug effects.

AVAILABILITY

The program and data used in this study is publically available at: https://github.com/aibotina/MCAAS.git.

AUTHOR CONTRIBUTIONS

PW developed the main method, designed and implemented the experiments, analyzed the result, and wrote the paper. WZ developed the method, and designed and implemented the experiments. BL conceived the project, and developed the method. LC conceived the project, and analyzed the result. LP designed and implemented the experiments, and analyzed the result. JY conceived the project, developed the main method, analyzed the result, and modified the paper.

FUNDING

This work was supported by National Nature Science Foundation of China (Grant Nos. 61863010, 61873076, 61370171, 61300128, 61472127, 61572178, 61672214, 61702054, 61803151, and 61772192) and the Natural Science Foundation of Hunan, China (Grant Nos. 2018JJ2461 and 2018JJ3568).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02500/full#supplementary-material

REFERENCES

Barnett, J. L., Yang, J., Cai, Z., Zhang, T., and Wan, X. F. (2012). AntigenMap 3D: an online antigenic cartography resource. *Bioinformatics* 28, 1292–1293. doi: 10.1093/bioinformatics/bts105

Barr Ig, E. A. (2014) Writing Committee of the World Health Organization Consultation on Northern Hemisphere Influenza Vaccine Composition for 2013–2014. WHO recommendations for the viruses used in the 2013-2014 Northern Hemisphere influenza vaccine: Epidemiology, antigenic and genetic characteristics of influenza A (H1N1)pdm09, A(H3N2) and B influenza viruses collected from October 2012 to January 2013. *Vaccine* 32, 4713–4725. doi: 10.1016/j.vaccine.2014.02.014

Blackburne, B. P., Hay, A. J., and Goldstein, R. A. (2008). Changing selective pressure during antigenic changes in human influenza H3. *Plos Pathogens* 4:e1000058. doi: 10.1371/journal.ppat.1000058

Cai, Z., Zhang, T., and Wan, X. F. (2010). A computational framework for influenza antigenic cartography. *Plos Computat. Biol.* 6:e1000949. doi: 10.1371/journal.pcbi.1000949

Chen, J., Li, K., Rong, H., Bilal, K., Yang, N., and Li, K. (2018). A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *Informationences*, 435, 124–149. doi: 10.1016/j.ins.2018.01.001

Fouchier, R. A. M., Smith, D. J., and Suarez, D. L. (2010). Use of antigenic cartography in vaccine seed strain selection. *Avian Diseases* 54, 220–223. doi: 10.1637/8740-032509-ResNote.1

Hay, A. J., Gregory, V., Douglas, A. R., and Lin, Y. P. (2001). The evolution of human influenza viruses. *Philos. Trans. R. Soc. Lond.* 356:1861. doi: 10.1098/rstb.2001.0999

Hirst, G. K. (1943). Studies of antigenic differences among strains of influenza a by means of read cell agglutination. *J. Exp. Med.* 78, 407–423. doi: 10.1084/jem.78.5.407

Huang, J. W., King, C. C., and Yang, J. M. (2009). Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. *BMC Bioinformat.* 10, 1–10. doi: 10.1186/1471-2105-10-S1-S41

Huang, L., Li, X., Guo, P., Yao, Y., Liao, B., Zhang, W., et al. (2017). Matrix completion with side information and its applications in predicting the antigenicity of influenza viruses. *Bioinformatics* 33:3195. doi: 10.1093/bioinformatics/btx390

Keshavan, R. H., Montanari A., and Sewoong, O. (2009a). Matrix completion from noisy entries. *J. Mach. Learn. Res.* 11, 2057–2078.

Keshavan, R. H., Oh, S., and Montanari, A. (2009b). Matrix completion from a few entries. *IEEE Int. Sympos. Informat. Theory.* 56, 2980–2998. doi: 10.1109/TIT.2010.2046205

Lapedes, A., and Farber, R. (2001). The geometry of shape space: application to influenza. *J. Theor. Biol.* 212:57. doi: 10.1006/jtbi.2001.2347

Lee, H., Lee, S., and Chang, H. L. (2016). Stochastic methods for epidemic models: an application to the 2009 H1N1 influenza outbreak in Korea. *Appl. Math. Comput.* 286, 232–249. doi: 10.1016/j.amc.2016.04.019

Lee, M. S., and Chen, J. S. (2004). Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg. Infect. Dis.* 10, 1385–1390. doi: 10.3201/eid1008.040107

Li, C., Tao, C., Qiang, H., Zhu, Y., and Li, K. (2016). Mruninovo : an efficient tool for de novo peptide sequencing utilizing the hadoop distributed computing framework. *Bioinformatics* 33, 944–946. doi: 10.1093/bioinformatics/btw721

Li, X., Liao, B., Cai, L., Cao, Z., and Zhu, W. (2013). Informative snps selection based on two-locus and multilocus linkage disequilibrium: criteria of max-correlation and min-redundancy. *IEEE. Trans. Comput. Biol. Bioinformat.* 10, 688–695. doi: 10.1109/TCBB.2013.61

Liang, Y., Qiu, K., Liao, B., Zhu, W., Huang, X., Li, L., et al. (2016). seeksv: an accurate tool for somatic structural variation and virus integration detection. *Bioinformatics* 33:184. doi: 10.1093/bioinformatics/btw591

Liao, B., Jiang, Y., Liang, W., Peng, L., Peng, L., Hanyurwimfura, D., et al. (2015a). On Efficient Feature Ranking Methods for High-Throughput Data Analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12:1374. doi: 10.1109/TCBB.2015.2415790

Liao, B., Li, X., Cai, L., Cao, Z., and Chen, H. (2015b). A hierarchical clustering method of selecting kernel SNP to unify informative SNP and tag SNP. *Comput. Biol. Bioinform. IEEE/ACM Trans.* 12, 113–122. doi: 10.1109/TCBB.2014.2351797

Liao, B., Li, X., Zhu, W., and Cao, Z. (2012b). A novel method to select informative snps and their application in genetic association studies. *IEEE Trans. Comput. Biol. Bioinformat.* 9:1529. doi: 10.1109/TCBB.2012.70

Liao, B., Li, X., Zhu, W., Li, R., and Wang, S. (2012a). Multiple ANT colony algorithm method for selecting tag SNPs. *J. Biomed. Informat.* 45:931. doi: 10.1016/j.jbi.2012.03.003

Liao, B., Liao, B., Sun, X., and Zeng, Q. (2010). A Novel method for similarity analysis and protein sub-cellular localization prediction. *Bioinformatics* 26, 2678–2683. doi: 10.1093/bioinformatics/btq521

Liao, Y. C., Lee, M. S., Ko, C. Y., and Chao, A. H. (2008). Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics* 24, 505–512. doi: 10.1093/bioinformatics/btm638

Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A., and Shraiman, B. I. (2016). Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc. Natl. Acad. Sci. U.S.A.* 113:E1701. doi: 10.1073/pnas.1525578113

Peng, L., Liao, B., Zhu, W., Li, Z., and Li, K. (2017). Predicting drug-target interactions with multi-information fusion. *IEEE J. Biomed. Health Inform.* 21, 561–572. doi: 10.1109/JBHI.2015.2513200

Shuichi Kawashima, P. Pokarowski, P., Pokarowska, P., Kolinski, A., Katayama, T., Kanehisa, A.M. (2008). Aaindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. doi: 10.1093/nar/gkm998

Smith, D. J., Lapedes, A. S., De Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D., et al. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science* 305, 371–376. doi: 10.1126/science.1097211

Sun, H., Yang, J., Zhang, T., Long, L. P., Jia, K., Yang, G., et al. (2013). Using sequence data to infer the antigenicity of influenza virus. *Mbio* 4, 234–250. doi: 10.1128/mBio.00230-13

Wang, W., Defeo, C. J., Alvaradofacundo, E., Vassell, R., and Weiss, C. D. (2015). Intermonomer Interactions in Hemagglutinin Subunits HA1 and HA2 Affecting Hemagglutinin Stability and Influenza Virus Infectivity. *J. Virol.* 89, 10602–10611. doi: 10.1128/JVI.00939-15

Wei, L., and Zou, Q. (2016). Recent progress in machine learning-based methods for protein fold recognition. *Int. J. Mol. Sci.* 17:2118. doi: 10.3390/ijms17122118

WHO (2017). *Influenza (Seasonal)*. Available online at: http://www.who.int/mediacentre/factsheets/fs211/en/ (Accessed)

Yang, J., Zhang, T., and Wan, X. F. (2014). Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. *PLoS ONE* 9:e106660. doi: 10.1371/journal.pone.0106660

Yao, Y., Li, X., Bo, L., Li, H., He, P., Wang, F., et al. (2017). Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method. *Sci. Rep.* 7:1545. doi: 10.1038/s41598-017-01699-z

Conflict of Interest Statement: 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.

Copyright © 2018 Wang, Zhu, Liao, Cai, Peng and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 October 2018
doi: 10.3389/fmicb.2018.02552

Dietary Exposure to the Environmental Chemical, PFOS on the Diversity of Gut Microbiota, Associated With the Development of Metabolic Syndrome

*Keng Po Lai[1], Alice Hoi-Man Ng[2], Hin Ting Wan[2], Aman Yi-Man Wong[2], Cherry Chi-Tim Leung[2], Rong Li[2] and Chris Kong-Chu Wong[2]**

[1] Department of Chemistry, City University of Hong Kong, Kowloon Tong, Hong Kong, [2] Croucher Institute for Environmental Sciences, Department of Biology, Hong Kong Baptist University, Kowloon Tong, Hong Kong

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Alinne Castro,
Universidade Católica Dom Bosco (UCDB), Brazil
Andrea Barbarossa,
Università degli Studi di Bologna, Italy

***Correspondence:**
Chris Kong-Chu Wong
ckcwong@hkbu.edu.hk

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

Received: *22 June 2018*
Accepted: *05 October 2018*
Published: *24 October 2018*

Citation:
Lai KP, Ng AH-M, Wan HT, Wong AY-M, Leung CC-T, Li R and Wong CK-C (2018) Dietary Exposure to the Environmental Chemical, PFOS on the Diversity of Gut Microbiota, Associated With the Development of Metabolic Syndrome. Front. Microbiol. 9:2552. doi: 10.3389/fmicb.2018.02552

The gut microbiome is a dynamic ecosystem formed by thousands of diverse bacterial species. This bacterial diversity is acquired early in life and shaped over time by a combination of multiple factors, including dietary exposure to distinct nutrients and xenobiotics. Alterations of the gut microbiota composition and associated metabolic activities in the gut are linked to various immune and metabolic diseases. The microbiota could potentially interact with xenobiotics in the gut environment as a result of their board enzymatic capacities and thereby affect the bioavailability and toxicity of the xenobiotics in enterohepatic circulation. Consequently, microbiome-xenobiotic interactions might affect host health. Here, we aimed to investigate the effects of dietary perfluorooctane sulfonic acid (PFOS) exposure on gut microbiota in adult mice and examine the induced changes in animal metabolic functions. In mice exposed to dietary PFOS for 7 weeks, body PFOS and lipid contents were measured, and to elucidate the effects of PFOS exposure, the metabolic functions of the animals were assessed using oral glucose-tolerance test and intraperitoneal insulin-tolerance and pyruvate-tolerance tests; moreover, on Day 50, cecal bacterial DNA was isolated and subject to 16S rDNA sequencing. Our results demonstrated that PFOS exposure caused metabolic disturbances in the animals, particularly in lipid and glucose metabolism, but did not substantially affect the diversity of gut bacterial species. However, marked modulations were detected in the abundance of metabolism-associated bacteria belonging to the phyla Firmicutes, Bacteroidetes, Proteobacteria, and Cyanobacteria, including, at different taxonomic levels, *Turicibacteraceae*, *Turicibacterales*, *Turicibacter*, *Dehalobacteriaceae*, *Dehalobacterium*, *Allobaculum*, *Bacteroides acidifaciens*, *Alphaproteobacteria*, and *4Cod-2/YS2*. The results of PICRUSt analysis further indicated that PFOS exposure perturbed gut metabolism, inducing notable changes in the metabolism of amino acids (arginine, proline, lysine), methane, and a short-chain fatty acid (butanoate), all of which

are metabolites widely recognized to be associated with inflammation and metabolic functions. Collectively, our study findings provide key information regarding the biological relevance of microbiome–xenobiotic interactions associated with the ecology of gut microbiota and animal energy metabolism.

Keywords: gut microbiome, bacterial diversity, microbiome-xenobiotic interaction, PFOs, energy metabolism

INTRODUCTION

The prevalence of non-communicable diseases (NCDs) is rapidly increasing, with cardiovascular diseases and diabetes being at the top of this list of NCDs, and multiple risk factors are widely recognized to be responsible for the increased incidences. Recently, the cumulative incidence of certain NCDs has been correlated to exposure to environmental chemicals (Gluckman et al., 2010). The past decades have witnessed the production of >150,000 synthetic chemicals, with approximately 2000 new chemicals being produced annually (Judson et al., 2009), and these heterogeneous chemical substances have been used for generating diverse industrial, agricultural, and commercial products. However, the release of these substances into the environment has adversely affected ecological and animal health. Depending on their chemical properties, these chemical substances have become dispersed in distinct environmental compartments and have contaminated food and water supplies. Retrospective analysis has revealed that exposure to various classes of environmental chemicals can occur through distinct routes and processes, including inhalation, dietary intake, and skin contact; this has resulted in the bodily accumulation of different environmental chemicals in the general population worldwide (Centers for Disease Control and Prevention, 2009), which indicates direct interactions of the exogenous chemicals within our body system.

The 2017 WHO global report on diabetes showed that >422 million adults were diagnosed with diabetes, underpinning the high prevalence of the disease and associated metabolic syndromes. People with a susceptible genetic background are predisposed to developing these diseases, and consumption of calorically dense diets and physical inactivity are the major risk factors associated with the disease development. However, these factors cannot account for the widespread prevalence of metabolic diseases in recent years, and thus additional investigation is required to reveal the pathogenesis of these diseases. Recently, scientific research has been focused on other potential risk factors that might disrupt body energy homeostasis, and considerable attention has been attracted by the roles of (1) gastrointestinal microbiota and (2) endocrine-disrupting chemicals (EDCs) as contributing factors.

The animal gut microbiome is a dynamic ecosystem formed by thousands of distinct bacterial species (Qin et al., 2010), and the remarkable metabolic activity in the gut environment is driven through a complex symbiotic interaction between these species. The gut bacterial diversity is shaped over time, with the complexity increasing due to the combined effect of multiple factors (such as genotype, diet composition, antibiotic therapy, and environmental exposure to xenobiotics) (Spor et al., 2011; Yatsunenko et al., 2012). The gut microbiota can potentially interact with environmental chemicals by altering the processes of absorption, disposition, metabolism, and excretion. Accordingly, gut bacteria have been widely reported to exhibit board ability to metabolize various environmental chemicals by using enzyme families (e.g., azoreductases, β-glucuronidases, β-lyases, nitroreductases, sulfatases) to catalyze diverse chemical reactions (e.g., reduction, hydrolysis, dehydroxylation, deacetylation, dinitration, deconjugation, demethylation) (Eriksson and Gustafsson, 1970; Williams et al., 1970; Bakke and Gustafsson, 1986; Rafii et al., 1990; Rafil et al., 1991; Roldan et al., 2008). A recent register of gut microbial biocatalytic reactions on xenobiotics listed 529 microorganisms that affect >1369 compounds (Gao et al., 2010); the study highlighted the capacity of gut microbes to transform diverse types of environmental chemicals. Notably, emerging evidence has indicated an association between body burden of environmental chemicals and gut microbial communities in the development of metabolic diseases (Alonso-Magdalena et al., 2011). In 2011, the U.S. National Toxicological Program studied the roles of environmental chemicals in the development of diabetes and obesity and reported positive correlations between EDC exposure and disease prevalence; the analysis prioritized the ten most predicted positive compounds across distinct biological processes: flusilazole, forchlorfenuron, d-cis/trans-allethrin, fentin, fludioxonil, niclosamide, prallethrin, thidiazuron, (Z,E)-fenpyroximate, and perfluorooctane sulfonic acid (PFOS). Among these chemicals, PFOS was listed as one of the risk factors for the development of metabolic diseases in the European research project OBELIX.

Alterations of gut microbiota composition are reported to be associated with various immune and metabolic diseases (e.g., inflammatory bowel disease, obesity, diabetes) (Cummings et al., 2003; Ley et al., 2006; Qin et al., 2012). However, few previous studies have investigated the interactions between environmental chemicals and gut microbiota and their toxicological relevance to the development of metabolic diseases. Here, we used a mouse model to assess the metabolic impact of dietary PFOS exposure. Physiological experiments and 16S rDNA metagenomic analyses were conducted to investigate the association among PFOS exposure, changes in gut bacterial community, and metabolic function.

MATERIALS AND METHODS

Experimental Animals and Chemicals

Female CD-1 mice (6–8 weeks old), obtained from the Animal Unit of the University of Hong Kong, were housed in

polypropylene cages containing sterilized bedding, maintained under a controlled temperature (23 ± 1°C, ambient temperature) and 12/12-h light/dark cycle, and provided *ad libitum* access to standard chow (LabDiet, 5001 Rodents Diet) and water (in glass bottles). The animal handling protocol was approved by the Committee on the Use of Human and Animal Subjects of the Hong Kong Baptist University (Permit no. 261812), in accordance with the Guidelines and Regulations of Department of Health, the Government of Hong Kong Special Administrative Region. The mice were acclimatized for 1 week before the PFOS-exposure experiments and then randomly divided into three groups (control, AC; low-dose PFOS, AL; high-dose PFOS, AH; at least four mice/group). PFOS (98% pure, Sigma-Aldrich) was dissolved in dimethyl sulfoxide (DMSO, Sigma-Aldrich) before mixing with corn oil; the final concentration of DMSO was <0.05% in all groups. The PFOS-exposure groups were weighed using an electronic balance (Shimadzu, Tokyo, Japan) and administered, every morning by oral gavage, 0.3 (AL) or 3 μg/g/day (AH) PFOS in corn oil for 7 weeks. The exposure doses were selected as described in our previous study (Lai et al., 2017), in reference with the human tolerable daily intake of PFOS established by the Scientific Panel on Contaminants in the Food Chain (European Food Safety Authority, 2008). The dose-range corresponded to the general population and occupational exposure levels. The control group (AC) received corn oil mixed with DMSO (0.05%).

Animals were sacrificed on Day 50 by cervical dislocation and cecal samples were collected. Blood samples were collected through cardiocentesis, and blood serum was prepared by centrifuging the samples at 3000 × *g* for 15 min. The serum and the weighed liver samples were stored at −20°C and then used for triglyceride (TG) and PFOS measurements.

Serum and Liver TGs

Serum and liver TG levels were quantified using the method described in our previous study (Wan et al., 2012) and a TG assay kit (Cayman, United States). Briefly, tissue samples were homogenized in chloroform:methanol (2:1) solution and then 0.05% sulfuric acid was added for phase separation. The aqueous phase was discarded, and the organic phase was collected and blow-dried under nitrogen gas at room temperature. The pellet was reconstituted in deionized water for 30 min at 37°C and then used for TG measurement.

Chemicals and Instrumental Analysis for PFOS

We used a mass-labeled mixed standard solution for perfluorinated compounds (Product code: MPFAC-MXA; Lot number: MPFACMXA0714; >98% pure) from Wellington Laboratories (ON, Canada). Samples were extracted and analyzed as previously described (Wan et al., 2013). Briefly, each tissue sample was mixed with 2 ng of internal standard, 1 mL of 0.5 M tetrabutylammonium hydroxide solution, 2 mL of 0.25 M sodium carbonate buffer, and 5 mL of methyl tert-butyl ether, and this was followed by mixing in a reciprocating shaker (HS 501 digital shaker, Janke and Kunkel IKA Labortechnik) at 250 rpm for 20 min. The organic and aqueous layers were separated, and the organic phase was collected, and the extraction procedure was repeated and all organic phases were pooled. The solution was blow-dried under nitrogen gas ($N_2 \geq 99.995\%$, Hong Kong Oxygen) in a nitrogen evaporator (N-EVAP112, Organomation Associates, Inc., MA, United States) and re-dissolved in 40% acetonitrile/60% 10 mM of ammonium acetate in Milli-Q water. An Agilent 1200 series liquid-chromatography system (Waldbronn, Germany) was used for PFOS detection. Chromatographic separation was performed using an Agilent ZORBAX Eclipse Plus C8 Narrow Bore guard column and an Agilent ZORBAX Eclipse Plus C8 Narrow Bore column. Tandem mass detection was conducted using an Agilent 6410B Triple Quadrupole mass spectrometer equipped with an Agilent Masshunter Workstation (version B.02.01) and an electrospray ionization source. The values of matrix recoveries were all 99%.

Physiological Analysis

Oral glucose-tolerance test (OGTT) and intraperitoneal (i.p.) insulin-tolerance test (ITT) and pyruvate-tolerance test (PTT) were conducted on Day 50 on control and PFOS-exposed mice, as described in our previous study (Wan et al., 2014). Briefly, for OGTT, 16-h-fasted mice were administered glucose (2 mg/g body weight); for PTT, 16-h-fasted mice received an i.p. injection of sodium pyruvate (2 mg/g of body weight); and for ITT, 12-h-fasted mice received an i.p. injection of insulin (1 IU/kg body weight). For measuring blood glucose, blood samples were collected by means of tail prick at 0, 15, 30, 60, and 120 min. Area under the curve (AUC) values for OGTT, PTT, and ITT were calculated to evaluate glucose tolerance, the total glucose synthesized from pyruvate, and insulin sensitivity, respectively.

16S rDNA Metagenomic Sequencing

Cecal bacterial DNA was isolated using DNeasy Blood & Tissue Kit (Qiagen), according to manufacturer instructions, and 30 ng of qualified DNA was used to construct the library for metagenomic sequencing. V3-V4 Dual-index Fusion PCR Primer Cocktail and PCR Master Mix were used to amplify the V3-V4 regions of 16S rDNA, and the PCR product was purified using Ampure XP beads (Agencourt). The library was quantified using real-time quantitative PCR and was quality-checked using an Agilent 2100 bioanalyzer instrument (EvaGreen). The normalized library was subject to Illumina MiSeq sequencing for 250-bp paired-end sequencing; sequencing data have been deposited in the NCBI Sequence Read Archive (SRA)[1], accession code SRP156864.

Bioinformatics Analysis

To obtain accurate and reliable results in bioinformatics analyses, we used a dual-indexing approach (Fadrosh et al., 2014). Raw data were filtered to eliminate adapters and low-quality reads by using an in-house procedure; this included truncation of sequencing reads, based on the phred algorithm: the removed sequencing reads (1) were <75% of their original length and contained their paired reads; (2) included adapter sequences (default

[1]http://www.ncbi.nlm.nih.gov/sra

parameter: 15 bases overlapped by reads); (3) contained an ambiguous base (N base) and their paired reads; and (4) exhibited low complexity (default: reads containing the same base in 10 consecutive positions). For pooling the library with barcoded samples, the clean reads were assigned to corresponding samples by allowing 0 base mismatch to barcode sequences with in-house scripts. The data-processing results are listed in **Supplementary Table S1**. At least 2 Mbp of clean data were obtained from each sample, and the read-usage ratio was >70%. Paired-end reads featuring overlaps were merged to tags that were clustered to operational taxonomic units (OTUs) by using the scripts of USEARCH software (v7.0.1090) (Edgar, 2013). All tags were clustered to OTUs at 97% sequence similarity. Taxonomic ranks were assigned to OTU representative sequences by using Ribosomal Database Project (RDP) Naïve Bayesian Classifier v.2.2. Alpha-diversity analysis and the screening for different species were based on OTU and taxonomic ranks. Phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt) analysis was employed to predict functional capabilities by using our sequencing data (Langille et al., 2013).

Statistical Analysis

Data are presented as means ± SEM. Differences between treatment and respective control groups were analyzed using Student's *t*-test; $p < 0.05$ was considered significant. Analyses were conducted using SigmaStat for Windows.

RESULTS

Effect of Chronic Dietary PFOS Exposure on Liver Weight and TG Content

Upon completion of the PFOS-exposure study, on Day 50, the mean body weights were increased in the control (AC) group and the low-dose (AL) and high-dose (AH) PFOS-treatment groups, and the weights did not differ in a statistically significant manner among the groups. However, in the AH group, the liver was enlarged and the absolute liver weight was increased (**Figure 1A**), as was the ratio of liver weight to body weight (**Figure 1B**). The liver appeared yellowish in the AH-group mice (data not shown), which might be associated with lipid accumulation. Accordingly, measurement of liver TG content revealed a significant increase in the PFOS-exposed mice (**Figure 1C**), and the liver TG content was positively correlated with the increase in absolute liver weight. Intriguingly, serum TG content in the AH group was significantly decreased relative to control (**Figure 1D**). **Table 1** shows the PFOS levels in both the liver and the serum in control and treatment groups; the accumulated PFOS levels were increased in a PFOS dose-dependent manner.

OGTT, ITT, and PTT

On Day 50, mice from the control and low- and high-dose PFOS-exposure groups were prepared for testing glucose metabolism and insulin function. OGTT results revealed that whereas the low-dose PFOS treatment did not significantly affect glucose tolerance (**Figure 2A**), the high-dose treatment elicited an earlier response in the reduction of blood glucose levels, at 15 and 30 min ($p < 0.05$), following glucose administration. The AUCs of OGTT were similar between the AC and AL groups, but the AUC of the AH group was significantly lower than that of the AC group (control).

Next, ITT-based measurement of body insulin sensitivity revealed that the mice in the AC and AL groups showed similar rate and extent of responses (**Figure 2B**), but in the AH-group mice, plasma glucose after insulin treatment was significantly lower than that in the control group. Accordingly, the AUC of the AH group was significantly lower than that of the AC group.

Lastly, to measure the effect of PFOS exposure on gluconeogenesis, pyruvate (a gluconeogenic substrate) was administrated and the rate of pyruvate conversion to glucose was measured. The PTT results indicated that pyruvate conversion in the AL and AH groups was significantly decreased relative to that in the AC group (**Figure 2C**).

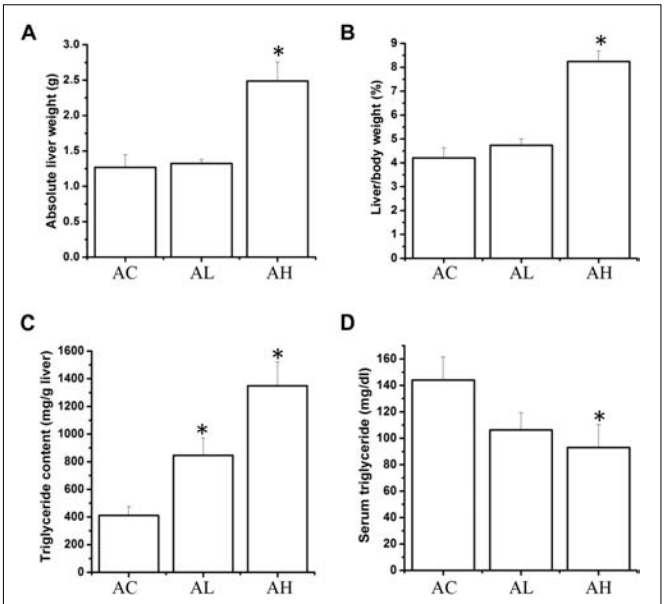


FIGURE 1 | Effect of 7-week dietary PFOS exposure on liver weight and triglyceride content in mice. **(A)** Absolute liver weight, **(B)** liver index, **(C)** liver triglyceride level, and **(D)** serum triglyceride level were measured on Day 50 after PFOS treatment. Data are presented as means ± SD; *$p < 0.05$ versus control group. AC, control; AL, 0.3 μg/g body weight/day; AH, 3 μg/g body weight/day.

TABLE 1 | Perfluorooctane sulfonic acid (PFOS) concentrations in liver and serum samples.

Sample name	PFOS (ng/g)	
	Liver	Serum
AC	7.10 ± 2.50	23.90 ± 8.18
AL	32942 ± 13473*	33781 ± 4365*
AH	503817 ± 325990*	109526 ± 19371*

Statistically significant, $p < 0.05$ as compared with the AC group. AC-control, AL-0.3 μg/g body weight per day, and AH-3 μg/g body weight per day.

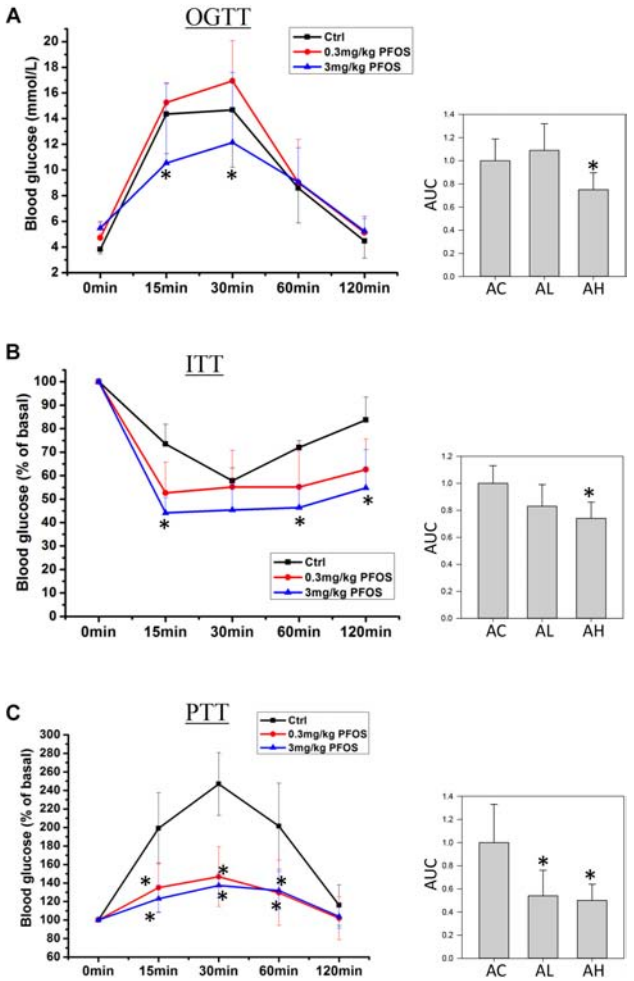


FIGURE 2 | Effect of 7-week dietary PFOS exposure in mice, examined using **(A)** oral glucose-tolerance test (OGTT), **(B)** insulin-tolerance test (ITT), and **(C)** pyruvate-tolerance test (PTT). Left panels: changes in serum glucose levels against time in the assays; right panels: area under curve (AUC) values of the respective assays. Data are presented as means ± SD; $*p < 0.05$ versus control group. AC, control; AL, 0.3 μg/g body weight/day; AH, 3 μg/g body weight/day.

PFOS Exposure Exerted No Marked Effect on Gut Bacterial Species Diversity

To determine the changes in gut bacterial community caused by chronic dietary PFOS exposure, we performed metagenomic sequencing analysis on the V3-V4 regions of 16S rDNA; the DNA was collected from the ceca of mice in the AC, AL, and AH groups. In OTU analysis, we found that the degree of bacterial diversity was similar among the groups, and the predominant phyla included Firmicutes, Bacteroidetes, and Proteobacteria. In the sample AH4, the number of bacterial species was low (**Table 2**). Venn diagram analysis (**Figure 3A**) revealed that 395 OTUs were shared among the three groups. Alpha diversity was next applied for analyzing the complexity of species diversity in each sample by using several indices: observed species, Chao1, ACE, Shannon, and Simpson indices (**Table 3**). The rarefaction curves based on the observed species value and Chao1 and ACE data were used to evaluate the coverage of the sequencing. The result showed that the sequencing data were adequate for covering all the bacterial species in the community, which was reflected in the appearance of plateau regions in the curves from all the samples (**Supplementary Figure S1**). Moreover, comparison of the species diversity in the three groups revealed that the PFOS-exposure groups showed no significant differences in gut bacterial species diversity relative to the control group (**Figure 3B**).

TABLE 2 | Operational taxonomic unit (OTU) analysis on each sample from the control (AC), low-dose (AL), and high-dose (AH) PFOS-exposed groups.

Sample name	Tag number	OTU number
AC1	67943	336
AC2	74582	301
AC3	74330	350
AC4	85158	284
AC5	88637	364
AH1	102682	351
AH2	91873	364
AH3	38690	255
AH4	75163	177
AL1	115877	385
AL2	55484	301
AL3	57234	318
AL4	57614	318
AL5	52056	306
AL6	71264	320

PFOS Exposure Altered Gut Microbiome Community at Different Taxonomic Levels

We compared the composition of the cecal microbiota at distinct taxonomic levels after dietary PFOS exposure. In the AL and AH groups, PFOS exposure produced similar and consistent effects in terms of changes in the abundance of certain microbial communities (**Table 4**). These changes included a significant increase at the level of the order *Turicibacterales* (belonging to the phylum Firmicutes) and a reduction of the species *Bacteroides acidifaciens* (phylum Bacteroidetes) (**Figure 4A** and **Table 4**); the increase in *Turicibacterales* was mainly contributed by an induction of the family *Turicibacteraceae* and genus *Turicibacter* (**Figure 4A** and **Table 4**). However, the abundance of certain other microbes was increased in either the AL group or the AH group: In the AL group, we identified a significant induction of the phylum Cyanobacteria (**Figure 4B** and **Table 4**), increases in *4Cod-2* (*Cyanobacteria*-like lineage) and the class *Alphaproteobacteria* (phylum Proteobacteria) (**Figure 4C** and **Table 4**), and an induction of the order YS2 (phylum Cyanobacteria) (**Figure 4D** and **Table 4**). Conversely, in the AH group, we detected a significant reduction in the family *Dehalobacteriaceae* (phylum Firmicutes) (**Figure 4E** and **Table 4**) and the genus *Dehalobacterium* (**Figure 4F** and

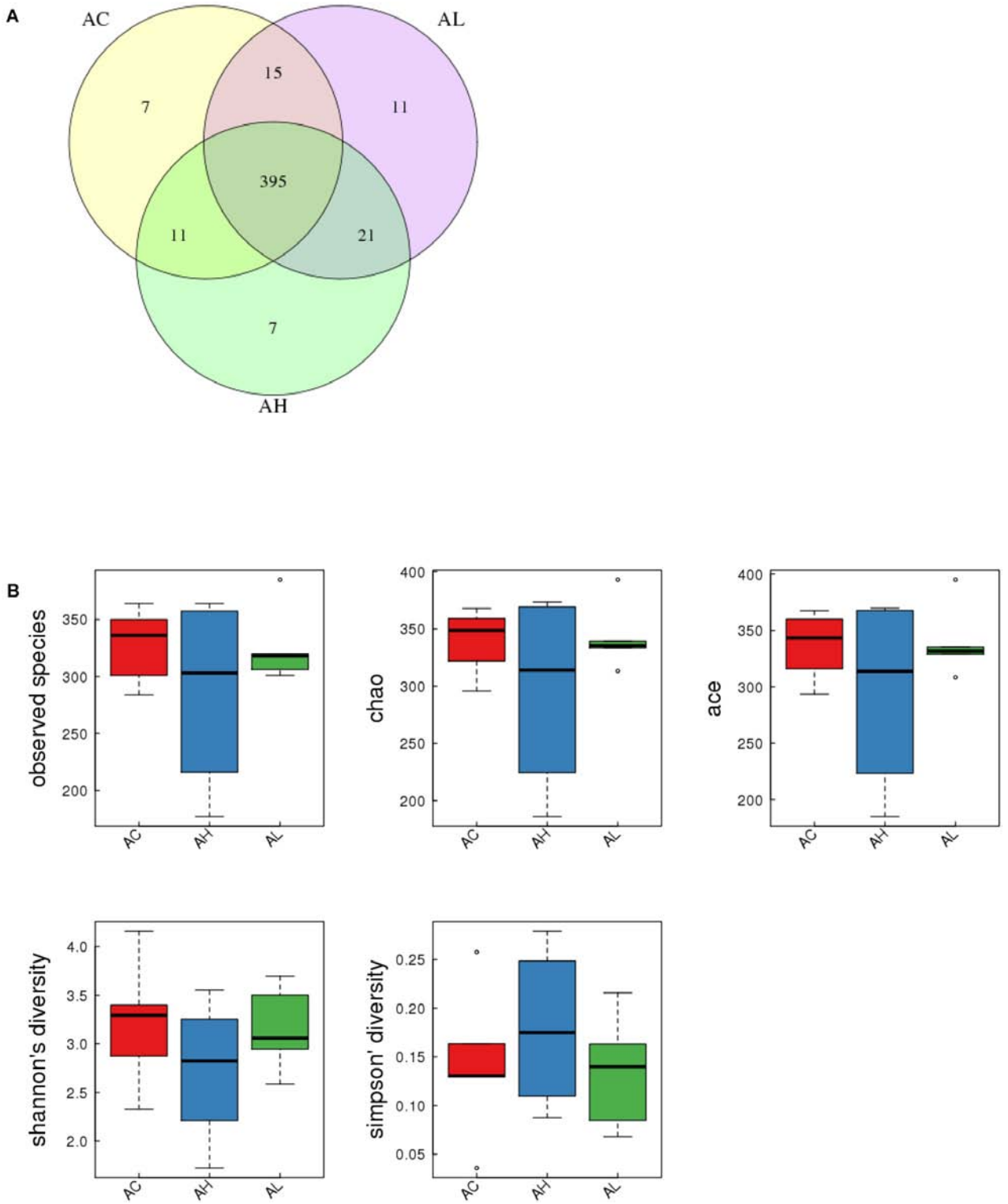


FIGURE 3 | Effect of 7-week dietary PFOS exposure on gut bacterial structure in mice. **(A)** Comparison of operational taxonomic units (OTUs); different colors represent distinct groups: (i) control (AC), (ii) low-dose PFOS exposure (AL), and (iii) high-dose PFOS exposure (AH). The intersection represents the set of OTUs commonly present in the counterpart groups. Venn diagram was drawn using VennDiagram software R (v3.0.3). **(B)** Changes in observed species number, Chao1 index, Ace index, Shannon's diversity, and Simpson's diversity; the results suggest that dietary PFOS intake exerted no effect on the species diversity of the gut bacterial community.

TABLE 3 | Alpha diversity statistics in each samples from the control (AC), low-dose (AL), and high-dose (AH) PFOS-exposed groups.

Sample name	Observed species	Chao1	ACE	Shannon	Simpson
AC1	336	348.75	343.26	4.16	0.04
AC2	301	322.00	315.92	2.33	0.26
AC3	350	359.13	360.02	3.40	0.13
AC4	284	295.67	293.48	2.87	0.16
AC5	364	367.93	367.49	3.29	0.13
AH1	351	365.00	365.72	2.96	0.13
AH2	364	373.55	369.73	3.55	0.09
AH3	255	263.00	261.75	2.69	0.22
AH4	177	186.07	184.91	1.72	0.28
AL1	385	393.08	395.01	3.50	0.08
AL2	301	313.21	308.38	3.14	0.12
AL3	318	339.43	329.40	2.95	0.16
AL4	318	336.07	328.85	3.70	0.07
AL5	306	333.36	335.19	2.59	0.22
AL6	320	334.77	333.86	2.97	0.16

TABLE 4 | Alteration of gut microbiome community at different taxonomy levels caused by dietary PFOS exposure.

Taxonomy level	Bacterial name	Ratio of low PFOS/normal diet	Ratio of high PFOS/normal diet
Phylum	*Cyanobacteria*	11.47*	4.98
Class	*Alphaproteobacteria*	19.25*	6.15
	4C0d-2	11.51*	5.01
Order	*Turicibacterales*	154.93*	68.49*
	YS2	11.51*	5.01
Family	*Turicibacteraceae*	154.93*	68.49*
	Dehalobacteriaceae	1.03	0.37*
Genus	*Dehalobacterium*	1.03	0.37*
	Allobaculum	9.82*	29.73
	Turicibacter	154.93*	68.49*
Species	*Bacteroides acidifaciens*	0.19*	0.23*

**Statistically significant change, $p < 0.05$.*

Table 4). Collectively, our results demonstrated that dietary PFOS exposure led to changes in the abundance of specific members of the gut-microbiome bacterial community.

PFOS Exposure Altered Gut Metabolism

Phylogenetic investigation of communities by reconstruction of unobserved states analysis was conducted to predict the functional profiling of gut bacterial communities in response to PFOS exposure. Our result demonstrated that both high- and low-dose PFOS exposure led to significant suppression of arginine and proline metabolism (**Table 5**). Moreover, high-dose PFOS exposure significantly reduced lysine biosynthesis and methane metabolism but induced butanoate metabolism. Taken together, these data suggest that PFOS exposure resulted in the alteration of gut metabolism.

DISCUSSION

Perfluorooctane sulfonic acid represents a risk factor for the development of metabolic diseases. A cross-sectional study conducted using data from the U.S. National Health and Nutrition Examination Survey 1999–2000 and 2003–2004, which examined 474 adolescents and 969 adults, reported that high plasma concentrations of PFOS were associated with increased blood insulin levels (Lin et al., 2009). In an evaluation of a potential link between plasma PFOS levels in 571 Taiwanese workers and the risk of diabetes, elevated levels of the chemical were correlated with impaired glucose homeostasis and increased prevalence of diabetes (Su et al., 2016). Furthermore, experimental studies in animal and cell models have demonstrated that PFOS exposure alters glucose and/or lipid metabolism through perturbations of pancreatic β-cells, adipocytes, and liver function, and in studies on adult-stage animals, chronic PFOS exposure has been found to reduce body weight and fat, accompanied by an increase in liver mass (Lau et al., 2007; Martin et al., 2007; Zhang et al., 2008; Cui et al., 2009). In the previous studies, most experiments were conducted using high-dose and acute PFOS exposure, and the experimental setting was thus unlike that in the real-world scenario, where low-dose and chronic exposure occurs. Moreover, limited information on the roles of gut microbes in PFOS-exposed animals is currently available. Therefore, our study was designed to address this knowledge gap. In the biochemical analysis of body TG content, our data revealed hepatomegaly and lipid accumulation in the liver of AH-group mice. The observation of liver enlargement and lipid accumulation agreed with the results of our previous study in which we used higher PFOS doses (5 and 10 μg/g body weight/day) but a shorter exposure time (21 days) (Wan et al., 2012). The hepatic lipid content might be increased because of the impairment of lipid catabolism and/or hepatic lipid export; the reduction in lipid catabolism probably occurred due to an inhibition of β-oxidation, whereas the reduction in lipid transport was related to a downregulation of apolipoprotein B (Wan et al., 2012). This correlation was further supported by the results obtained in this study, which showed a marked reduction in serum TG level in the AH group. The perturbation of lipid metabolism could have affected glucose metabolism and insulin secretion (Antinozzi et al., 1998). Thus, we conducted physiological tests to evaluate the impact of PFOS exposure on glucose tolerance, insulin sensitivity, and hepatic gluconeogenesis. Our results showed statistically significant changes in the responses measured in OGTT and ITT in the AH group. The findings of both assays suggested that the high-dose PFOS exposure induced insulin hypersensitivity in mice, with the evidence indicating an increased rate of reduction of plasma glucose levels and a decreased rate of gluconeogenesis. The observation is supported by a previous study showing that exposure of mice to PFOA (perfluorooctanoic acid, a member of the PFOS family) led to an elevation of insulin sensitivity (Yan et al., 2015). One of the recognized physiological functions of insulin is to promote hepatic fatty acid synthesis. The high liver lipid content in the AH group appeared to be the biological outcome of this effect. Dietary PFOS exposure would lead to

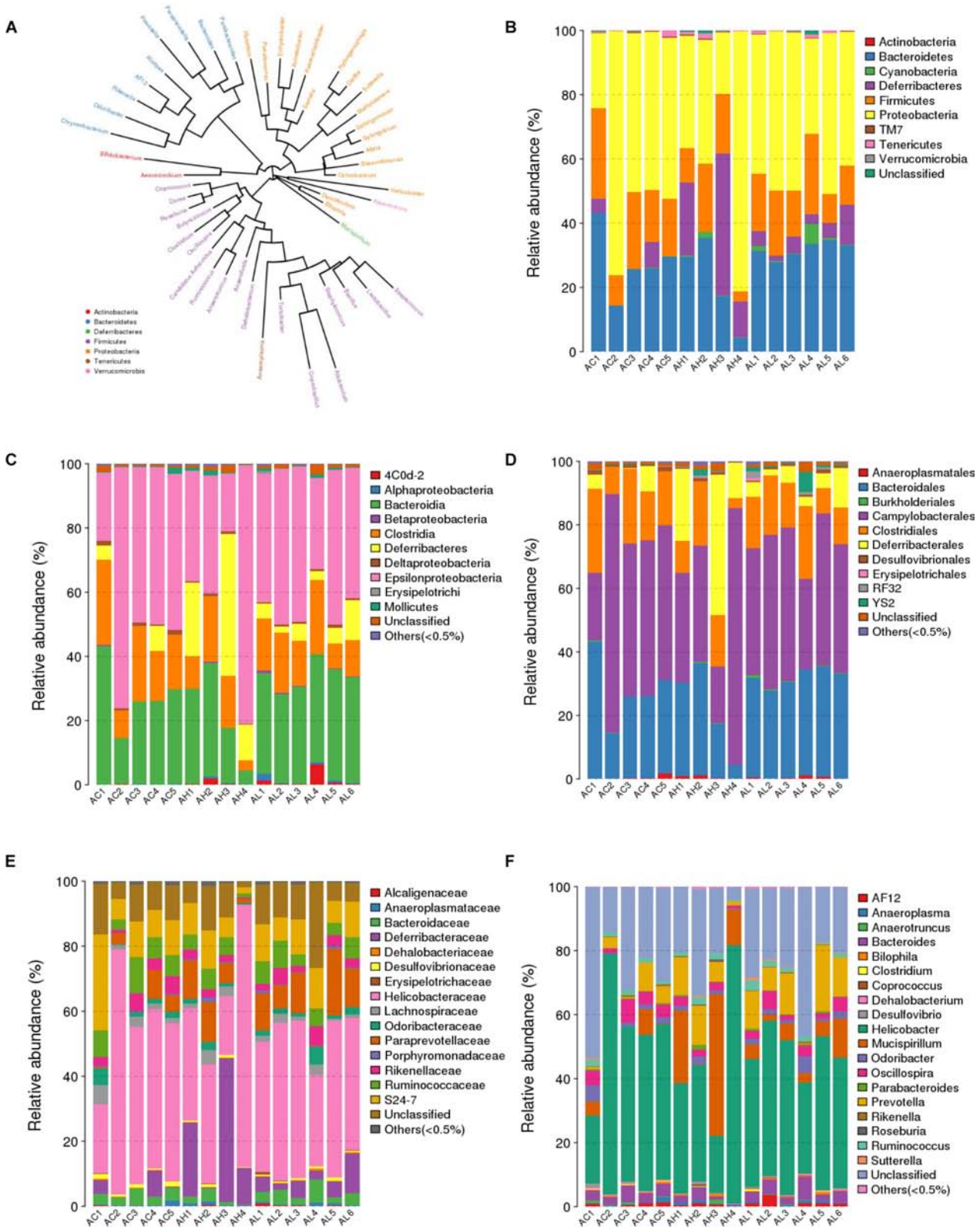


FIGURE 4 | Effect of 7-week dietary PFOS exposure on gut microbiome community at distinct taxonomic levels. **(A)** Phylogenetic tree diagram at genus level. The same color indicates the same phylum. Taxonomic composition distributions in control (AC), low-dose PFOS-exposure (AL), and high-dose PFOS-exposure (AH) groups are shown at the levels of **(B)** phylum, **(C)** class, **(D)** order, **(E)** family, and **(F)** genus.

TABLE 5 | Alteration of gut metabolisms caused by dietary PFOS exposure.

Functional classification	Ratio of low PFOS/normal diet	Ratio of high PFOS/normal diet
Arginine and proline metabolism	0.92*	0.93*
Butanoate metabolism	1.04	1.06*
Lysine biosynthesis	0.95	0.95*
Methane metabolism	0.91	0.87*

**Statistically significant change, $p < 0.05$.*

direct interaction of the chemical with the bacteria in the gut environment, and, intriguingly, this physiological outcome correlated with the changes in gut bacterial diversity assessed using 16S metagenomic analysis.

In our previous study in mice, we showed that daily intake of an environmental obesogen, bisphenol A, altered the gut bacterial structure (Lai et al., 2016). The pattern of the alteration was similar to that in high-fat-diet-fed mice. This observation supports the notion that environmental chemicals can perturb gut bacterial communities. In this study, we extended our investigation to address the effects of PFOS exposure on gut bacterial structure. Our results showed that chronic PFOS exposure (0.3 and 3 μg/g body weight, for 49 days) exerted no effect on gut bacterial diversity in general. However, when we examined specific taxonomic levels, we found that both low-dose and high-dose of PFOS exposure altered the abundances of distinct gut bacteria belonging to the phyla Firmicutes, Bacteroidetes, Proteobacteria, and Cyanobacteria. Some of these changes were reported to be associated with the symptoms of metabolic perturbations. For instance, PFOS exposure caused a marked induction of microbes in the order *Turicibacterales*, which was due to the growth of the bacteria in the family *Turicibacteraceae* and genus *Turicibacter*, and this induction was stronger in the low-dose PFOS-exposure group than in the high-dose group. A previous study showed that *Turicibacter* was increased in mice fed with a high-cholesterol diet, as compared with the level in the control group (Dimova et al., 2017); the data implied that *Turicibacter* was increased in response to the abundance of dietary cholesterol. Intriguingly, the results of an epidemiological analysis showed a positive correlation between serum PFOS and total cholesterol levels (Nelson et al., 2010). Moreover, other studies suggested that an increase in *Turicibacter* was correlated with dietary fat content, although the observations were inconclusive (Everard et al., 2014; Zhong et al., 2015). Nonetheless, the increase we observed here in the abundance of *Turicibacter* was likely related to the perturbing effects of PFOS on lipid metabolism. Another study on host–microbiota relationship in glucose-metabolism disorder demonstrated a positive association with *Turicibacteraceae* (Lippert et al., 2017). This association was observed here in our OGTT, ITT, and PTT data, particularly in the case of high-dose PFOS exposure. Moreover, following low-dose PFOS exposure, the abundance of the genus *Allobaculum* was increased substantially. *Allobaculum*, a putative short-chain fatty-acid-producing bacterium, was suggested to contribute to insulin resistance and obesity (Zhang et al., 2015). Besides this increase of bacterial abundance, our data revealed a marked reduction in the proportion of *B. acidifaciens* in the gut of mice in the PFOS-exposed groups, as compared with the proportion in the control group. *B. acidifaciens* is one of the predominant bacterial species responsible for promoting IgA production in the large intestine and is a specific commensal bacterium associated with amelioration of metabolic disorders in mice (Yanagibashi et al., 2013; Yang et al., 2017). The abundance of *B. acidifaciens* was found to be negatively correlated with liver TG levels in mice fed with a high-fat diet (Blasco-Baque et al., 2017), which supports our data indicating negative correlation between the levels of hepatic and serum TG in PFOS-exposed mice. The family *Dehalobacteriaceae* showed reduced abundance only in the high-dose group, which was contributed by the decrease in *Dehalobacterium*. In a study of 416 twin-pairs from the Twins population, a low abundance of *Dehalobacterium* was associated with a high body mass index and high blood lipid levels (Fu et al., 2015). The involvement of the gut microbiota in multiple metabolic pathways in the host is widely recognized, and, accordingly, the results of our PICRUSt analysis showed that PFOS exposure altered the microbial community functions, specifically in the metabolism of amino acids (arginine, proline, lysine), methane, and a short-chain fatty acid (butanoate). Alternations in the metabolism of these metabolites in intestinal bacteria were reported to affect host physiology (Dai et al., 2011; Nicholson et al., 2012); changes in arginine and proline metabolism were associated with coronary heart disease (Feng et al., 2016), whereas perturbations of butyrate and methane metabolism were related to inflammatory diseases (Morgan et al., 2012) and Type I diabetes (Brown et al., 2011). Furthermore, the GPR-43 receptor for short-chain fatty acids was demonstrated to be linked with fat accumulation in the host (Kimura et al., 2013). Retrospectively, we can conclude that our data on the changes in the abundance of gut bacteria and their metabolism in the PFOS-exposed groups were associated with the observed metabolic perturbations.

To our knowledge, this the first integrative study to report the effects of PFOS exposure on animal metabolism and gut bacterial community. Our data revealed that chronic PFOS exposure at 3 μg/g body weight/day induced insulin sensitivity, which was associated with an increase in hepatic lipid content but a reduction in hepatic gluconeogenesis. The results of intestinal 16S metagenomic analysis demonstrated marked changes in the abundances of bacteria at distinct taxonomic levels, including *Turicibacter*, *Allobaculum*, *B. acidifaciens*, and *Dehalobacteriaceae*; changes in the abundance of these bacteria are known to be associated with perturbations of glucose and lipid metabolism. Collectively, the results from this study implied that dietary PFOS exposure affected not only the glucose and lipid metabolism of the host animals, but also caused disturbance to the gut bacterial ecosystem. However, certain questions remain unresolved, such as the mechanistic interactions between PFOS and gut microbes and the changes in the production of bacterial metabolites, and further investigation in necessary to clarify the potential correlation between these changes and PFOS exposure.

DATA AVAILABILITY

Sequence data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra); accession code: SRP156864.

AUTHOR CONTRIBUTIONS

KL participated in metagenomic sequencings, analyzed the data, and drafted the manuscript. HW, CL, and RL carried out the animal works and sample preparation. AW was involved in chemical analysis. AN carried out the bioinformatic data analysis. CW conceived the idea, formulated the hypothesis, and drafted the manuscript.

FUNDING

This work was supported by the Strategic Research Fund to CW (Hong Kong Baptist University) (RC-ICRS/17-18/01).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02552/full#supplementary-material

FIGURE S1 | Rarefaction curves based on Chao1 index, Ace index, and observed species values, showing that the data volume covered all species in the gut bacterial community.

TABLE S1 | The sequencing statistic of metagenomics sequencing.

REFERENCES

Alonso-Magdalena, P., Quesada, I., and Nadal, A. (2011). Endocrine disruptors in the etiology of type 2 diabetes mellitus. *Nat. Rev. Endocrinol.* 7, 346–353. doi: 10.1038/nrendo.2011.56

Antinozzi, P. A., Segall, L., Prentki, M., McGarry, J. D., and Newgard, C. B. (1998). Molecular or pharmacologic perturbation of the link between glucose and lipid metabolism is without effect on glucose-stimulated insulin secretion. A re-evaluation of the long-chain acyl-CoA hypothesis. *J. Biol. Chem.* 273, 16146–16154. doi: 10.1074/jbc.273.26.16146

Bakke, J. E., and Gustafsson, J. A. (1986). Role of intestinal flora in metabolism of agrochemicals conjugated with glutathione. *Xenobiotica* 16, 1047–1056. doi: 10.3109/00498258609038982

Blasco-Baque, V., Coupe, B., Fabre, A., Handgraaf, S., Gourdy, P., Arnal, J. F., et al. (2017). Associations between hepatic miRNA expression, liver triacylglycerols and gut microbiota during metabolic adaptation to high-fat diet in mice. *Diabetologia* 60, 690–700. doi: 10.1007/s00125-017-4209-3

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

Centers for Disease Control and Prevention. (2009). *Fourth National Report on Human Exposure to Environmental Chemicals, Department of Health and Human Services*. Atlanta, GA: Centers for Disease Control and Prevention.

Cui, L., Zhou, Q. F., Liao, C. Y., Fu, J. J., and Jiang, G. B. (2009). Studies on the toxicological effects of PFOA and PFOS on rats using histological observation and chemical analysis. *Arch. Environ. Contam. Toxicol.* 56, 338–349. doi: 10.1007/s00244-008-9194-6

Cummings, J. H., Macfarlane, G. T., and Macfarlane, S. (2003). Intestinal bacteria and ulcerative colitis. *Curr. Issues Intest. Microbiol.* 4, 9–20.

Dai, Z. L., Wu, G., and Zhu, W. Y. (2011). Amino acid metabolism in intestinal bacteria: links between gut ecology and host health. *Front. Biosci.* 16, 1768–1786. doi: 10.2741/3820

Dimova, L. G., Zlatkov, N., Verkade, H. J., Uhlin, B. E., and Tietge, U. J. F. (2017). High-cholesterol diet does not alter gut microbiota composition in mice. *Nutr. Metab.* 14:15. doi: 10.1186/s12986-017-0170-x

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Eriksson, H., and Gustafsson, J. A. (1970). Steroids in germfree and conventional rats. Sulpho- and glucuronohydrolase activities of caecal contents from conventional rats. *Eur. J. Biochem.* 13, 198–202. doi: 10.1111/j.1432-1033.1970.tb00919.x

European Food Safety Authority. (2008). Perfluorooctane sulfonate (PFOS), perfluorooctanic acid (PFOA) and their salts - scientific opinion of the panel on contaminants in the food chain. *EFSA J.* 653, 1–131.

Everard, A., Lazarevic, V., Gaia, N., Johansson, M., Stahlman, M., Backhed, F., et al. (2014). Microbiome of prebiotic-treated mice reveals novel targets involved in host response during obesity. *ISME J.* 8, 2116–2130. doi: 10.1038/ismej.2014.45

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., et al. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* 2:6. doi: 10.1186/2049-2618-2-6

Feng, Q., Liu, Z., Zhong, S., Li, R., Xia, H., Jie, Z., et al. (2016). Integrated metabolomics and metagenomics analysis of plasma and urine identified microbial metabolites associated with coronary heart disease. *Sci. Rep.* 6:22525. doi: 10.1038/srep22525

Fu, J., Bonder, M. J., Cenit, M. C., Tigchelaar, E. F., Maatman, A., Dekens, J. A., et al. (2015). The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circ. Res.* 117, 817–824. doi: 10.1161/CIRCRESAHA.115.306807

Gao, J., Ellis, L. B., and Wackett, L. P. (2010). The university of minnesota biocatalysis/biodegradation database: improving public access. *Nucleic Acids Res.* 38, D488–D491. doi: 10.1093/nar/gkp771

Gluckman, P. D., Hanson, M. A., and Mitchell, M. D. (2010). Developmental origins of health and disease: reducing the burden of chronic disease in the next generation. *Genome Med.* 2:14. doi: 10.1186/gm135

Judson, R., Richard, A., Dix, D. J., Houck, K., Martin, M., Kavlock, R., et al. (2009). The toxicity data landscape for environmental chemicals. *Environ. Health Perspect.* 117, 685–695. doi: 10.1289/ehp.0800168

Kimura, I., Ozawa, K., Inoue, D., Imamura, T., Kimura, K., Maeda, T., et al. (2013). The gut microbiota suppresses insulin-mediated fat accumulation via the short-chain fatty acid receptor GPR43. *Nat. Commun.* 4:1829. doi: 10.1038/ncomms2852

Lai, K. P., Chung, Y. T., Li, R., Wan, H. T., and Wong, C. K. (2016). Bisphenol A alters gut microbiome: comparative metagenomics analysis. *Environ. Pollut.* 218, 923–930. doi: 10.1016/j.envpol.2016.08.039

Lai, K. P., Lee, J. C., Wan, H. T., Li, J. W., Wong, A. Y., Chan, T. F., et al. (2017). Effects of in utero PFOS exposure on transcriptome, lipidome, and function of mouse testis. *Environ. Sci. Technol.* 51, 8782–8794. doi: 10.1021/acs.est.7b02102

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

Lau, C., Anitole, K., Hodes, C., Lai, D., Pfahles-Hutchens, A., and Seed, J. (2007). Perfluoroalkyl acids: a review of monitoring and toxicological findings. *Toxicol. Sci.* 99, 366–394. doi: 10.1093/toxsci/kfm128

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

Lin, C. Y., Chen, P. C., Lin, Y. C., and Lin, L. Y. (2009). Association among serum perfluoroalkyl chemicals, glucose homeostasis, and metabolic syndrome in adolescents and adults. *Diabetes Care* 32, 702–707. doi: 10.2337/dc08-1816

Lippert, K., Kedenko, L., Antonielli, L., Kedenko, I., Gemeier, C., Leitner, M., et al. (2017). Gut microbiota dysbiosis associated with glucose metabolism disorders and the metabolic syndrome in older adults. *Benef. Microbes.* 8, 545–556. doi: 10.3920/BM2016.0184

Martin, M. T., Brennan, R. J., Hu, W., Ayanoglu, E., Lau, C., Ren, H., et al. (2007). Toxicogenomic study of triazole fungicides and perfluoroalkyl acids in rat livers predicts toxicity and categorizes chemicals based on mechanisms of toxicity. *Toxicol. Sci.* 97, 595–613. doi: 10.1093/toxsci/kfm065

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. doi: 10.1186/gb-2012-13-9-r79

Nelson, J. W., Hatch, E. E., and Webster, T. F. (2010). Exposure to polyfluoroalkyl chemicals and cholesterol, body weight, and insulin resistance in the general U.S. population. *Environ. Health Perspect.* 118, 197–202. doi: 10.1289/ehp.0901165

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., et al. (2012). Host-gut microbiota metabolic interactions. *Science* 336, 1262–1267. doi: 10.1126/science.1223813

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

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. doi: 10.1038/nature11450

Rafii, F., Franklin, W., and Cerniglia, C. E. (1990). Azoreductase activity of anaerobic bacteria isolated from human intestinal microflora. *Appl. Environ. Microbiol.* 56, 2146–2151.

Rafil, F., Franklin, W., Heflich, R. H., and Cerniglia, C. E. (1991). Reduction of nitroaromatic compounds by anaerobic bacteria isolated from the human gastrointestinal tract. *Appl. Environ. Microbiol.* 57, 962–968.

Roldan, M. D., Perez-Reinado, E., Castillo, F., and Moreno-Vivian, C. (2008). Reduction of polynitroaromatic compounds: the bacterial nitroreductases. *FEMS Microbiol. Rev.* 32, 474–500. doi: 10.1111/j.1574-6976.2008.00107.x

Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* 9, 279–290. doi: 10.1038/nrmicro2540

Su, T. C., Kuo, C. C., Hwang, J. J., Lien, G. W., Chen, M. F., and Chen, P. C. (2016). Serum perfluorinated chemicals, glucose homeostasis and the risk of diabetes in working-aged Taiwanese adults. *Environ. Int.* 88, 15–22. doi: 10.1016/j.envint.2015.11.016

Wan, H. T., Leung, P. Y., Zhao, Y. G., Wei, X., Wong, M. H., and Wong, C. K. (2013). Blood plasma concentrations of endocrine disrupting chemicals in Hong Kong populations. *J. Hazard. Mater.* 261, 763–769. doi: 10.1016/j.jhazmat.2013.01.034

Wan, H. T., Zhao, Y. G., Leung, P. Y., and Wong, C. K. (2014). Perinatal exposure to perfluorooctane sulfonate affects glucose metabolism in adult offspring. *PLoS.One.* 9:e87137. doi: 10.1371/journal.pone.0087137

Wan, H. T., Zhao, Y. G., Wei, X., Hui, K. Y., Giesy, J. P., and Wong, C. K. (2012). PFOS-induced hepatic steatosis, the mechanistic actions on beta-oxidation and lipid transport. *Biochim. Biophys. Acta* 1820, 1092–1101. doi: 10.1016/j.bbagen.2012.03.010

Williams, J. R. Jr., Grantham, P. H., Marsh, H. H. III, Weisburger, J. H., and Weisburger, E. K. (1970). Participation of liver fractions and of intestinal bacteria in the metabolism of N-hydroxy-N-2-fluorenylacetamide in the rat. *Biochem. Pharmacol.* 19, 173–188. doi: 10.1016/0006-2952(70)90338-2

Yan, S., Zhang, H., Zheng, F., Sheng, N., Guo, X., and Dai, J. (2015). Perfluorooctanoic acid exposure for 28 days affects glucose homeostasis and induces insulin hypersensitivity in mice. *Sci. Rep.* 5:11029. doi: 10.1038/srep11029

Yanagibashi, T., Hosono, A., Oyama, A., Tsuda, M., Suzuki, A., Hachimura, S., et al. (2013). IgA production in the large intestine is modulated by a different mechanism than in the small intestine: Bacteroides acidifaciens promotes IgA production in the large intestine by inducing germinal center formation and increasing the number of IgA + B cells. *Immunobiology* 218, 645–651. doi: 10.1016/j.imbio.2012.07.033

Yang, J. Y., Lee, Y. S., Kim, Y., Lee, S. H., Ryu, S., Fukuda, S., et al. (2017). Gut commensal Bacteroides acidifaciens prevents obesity and improves insulin sensitivity in mice. *Mucosal. Immunol.* 10, 104–116. doi: 10.1038/mi.2016.42

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, 222–227. doi: 10.1038/nature11053

Zhang, H., Shi, Z., Liu, Y., Wei, Y., and Dai, J. (2008). Lipid homeostasis and oxidative stress in the liver of male rats exposed to perfluorododecanoic acid. *Toxicol. Appl. Pharmacol.* 227, 16–25. doi: 10.1016/j.taap.2007.09.026

Zhang, X., Zhao, Y., Xu, J., Xue, Z., Zhang, M., Pang, X., et al. (2015). Modulation of gut microbiota by berberine and metformin during the treatment of high-fat diet-induced obesity in rats. *Sci. Rep.* 5:14405. doi: 10.1038/srep14405

Zhong, Y., Nyman, M., and Fak, F. (2015). Modulation of gut microbiota in rats fed high-fat diets by processing whole-grain barley to barley malt. *Mol. Nutr. Food Res.* 59, 2066–2076. doi: 10.1002/mnfr.201500187

Conflict of Interest Statement: 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.

Copyright © 2018 Lai, Ng, Wan, Wong, Leung, Li and Wong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 26 October 2018
doi: 10.3389/fmicb.2018.02571

PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method

Yi Xiong[1], Qiankun Wang[1], Junchen Yang[1], Xiaolei Zhu[2] and Dong-Qing Wei[1]**

[1] State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [2] School of Sciences, Anhui Agricultural University, Hefei, China

Gram-negative bacteria use various secretion systems to deliver their secreted effectors. Among them, type IV secretion system exists widely in a variety of bacterial species, and secretes type IV secreted effectors (T4SEs), which play vital roles in host-pathogen interactions. However, experimental approaches to identify T4SEs are time- and resource-consuming. In the present study, we aim to develop an *in silico* stacked ensemble method to predict whether a protein is an effector of type IV secretion system or not based on its sequence information. The protein sequences were encoded by the feature of position specific scoring matrix (PSSM)-composition by summing rows that correspond to the same amino acid residues in PSSM profiles. Based on the PSSM-composition features, we develop a stacked ensemble model PredT4SE-Stack to predict T4SEs, which utilized an ensemble of base-classifiers implemented by various machine learning algorithms, such as support vector machine, gradient boosting machine, and extremely randomized trees, to generate outputs for the meta-classifier in the classification system. Our results demonstrated that the framework of PredT4SE-Stack was a feasible and effective way to accurately identify T4SEs based on protein sequence information. The datasets and source code of PredT4SE-Stack are freely available at http://xbioinfo.sjtu.edu.cn/PredT4SE_Stack/index.php.

Keywords: type IV secreted effector, sequence information, position specific scoring matrix, machine learning, stacked ensemble method

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Quan Zou,
Tianjin University, China
Zhenhua Li,
National University of Singapore, Singapore

****Correspondence:***
Xiaolei Zhu
xlzhu_mdl@hotmail.com
Dong-Qing Wei
dqwei@sjtu.edu.cn

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

Received: *01 September 2018*
Accepted: *09 October 2018*
Published: *26 October 2018*

Citation:
Xiong Y, Wang Q, Yang J, Zhu X and Wei D-Q (2018) PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method. Front. Microbiol. 9:2571. doi: 10.3389/fmicb.2018.02571

INTRODUCTION

Gram-negative bacteria use various secretion systems to deliver their secreted substrates (also called as effectors) from the bacterial cytosol into host cells, which can promote virulence and cause diseases. Until now, eight different secretion systems (type I to type VIII) have been found in Gram-negative bacteria, which differ from each other in their outer membrane secretion mechanisms. There are a number of well-organized databases or web resource on collecting experimentally validated effectors of Type III, IV, and VI secretion systems (Bi et al., 2013; Li et al., 2015; Eichinger et al., 2016; An et al., 2017). Among them, type IV secretion system (T4SS) exists widely in a variety of bacterial species, such as *Bordetella pertussis*, *Helicobacter pylori*, *Coxiella burnetii*, and *Legionella pneumophila* (Chandran et al., 2009; Fronzes et al., 2009; Lifshitz et al., 2013). T4SS

specifically secretes type IV secreted effectors (T4SEs), which vary widely across bacterial species. T4SEs mimic the function of host proteins, exert vital functions in cytoplasm of infected eukaryotic cells and play crucial roles in host-pathogen interactions. Accurate and reliable identification of T4SEs is a crucial step toward the understanding of the pathogenic mechanism of T4SS. Due to the biological significance of T4SEs, a number of experimental approaches have been developed to identify novel T4SEs such as fusion protein report assays and secretion apparatus. However, these experimental approaches are time- and resource-consuming. It is highly desirable to develop *in silico* classification models to accurately predict type IV secreted effectors of T4SS based on protein sequence information.

In the last decade, several computational approaches using machine learning (ML) algorithms were developed to predict T4SEs based on protein sequence information. A pioneering method proposed by Burstein et al. (2009) formulated the task of identifying T4SEs on Legionella pneumophila genome as a classification problem using various ML algorithms, including naïve Bayes, Bayesian networks, support vector machine (SVM), Neural networks, and a voting algorithm that is based on these four algorithms. The input features of these algorithms include taxonomical dispersion, regulatory data, genomic organization, and similarity to eukaryotic proteomes (Burstein et al., 2009). Later, the same group developed a hidden semi-Markov model (HSMM) to characterize the amino acid composition of the secretion signal for identification of T4SEs across species (Lifshitz et al., 2013). Chen et al. (2010) used the similar ML-based model as the previous study (Burstein et al., 2009) to predict putative T4SEs in Coxiella burnetii genome, which helped narrow the number of potential targets for subsequent experimental validation. T4EffPred is a SVM-based prediction tool for identifying T4SEs based on four types of sequence-derived features, which were calculated from amino acid composition (AAC) and position specific scoring matrix (PSSM) profiles (Zou et al., 2013). T4SEpre (Wang et al., 2014) is another SVM-based tool for predicting T4SEs from C-terminal 100 amino acids of protein sequences by using AAC, position-specific AAC profiles, and predicted structural features such as secondary structure and solvent accessibility. An et al. (2016) constructed an ensemble model by random forest to integrate the output of the individual predictors (i.e., T4EffPred and T4SEpre) to improve predictive performance. Recently, Wang Y. et al. (2017) presented an effective method to predict T4SEs prediction by integrating information from both 50 N-terminal and 100 C-terminal residues of protein sequences. The model was built by SVM based on three types of features, namely AAC, PSSM, and composition, transition and distribution.

Overall, the currently available computational approaches for prediction of T4SEs vary from one another in terms of the utilized features and ML algorithms. Since the numbers of effectors and non-effectors in genomes are heavily unbalanced (the effectors comprise only a small fraction of a genome), it is highly desirable to develop a prediction method with high precision and high specificity. Otherwise, the number of true positives would easily be overwhelmed by the number of false positives, so that such a predictor is impractical to generate reliable candidates for experimental validation. In the present study, we aim to propose a stacked ensemble model, PredT4SE-Stack, to further improve the prediction performance (i.e., higher precision and specificity) for identifying T4SEs from protein sequence information. The stacked generalization approach (Wolpert, 1992) consists of an ensemble of base classifiers whose outputs are further learned by a meta-classifier to model the relationship between the ensemble outputs and the actual classes/labels. To construct the model, the protein sequences are firstly encoded by the feature of PSSM-composition by summing rows that correspond to the same amino acid residues in PSSM profiles. Based on the PSSM-composition features, a total of eight types of ML-based algorithms (including advanced ML techniques) are used to build base-classifiers in the first stage. Then, the optimal combination of base-classifiers is searched, and the output of these selected base-classifiers are utilized as input for a meta-classifier at the second stage. Our experimental results on both cross validation and independent tests demonstrated that the framework of PredT4SE-Stack is a feasible and effective way to accurately identify T4SEs based on protein sequence information. It also has achieved better performance than previously published methods.

MATERIALS AND METHODS

Dataset

In this study, the same benchmark dataset curated by Wang Y. et al. (2017) was used to evaluate the performance of our proposed method. The dataset consists of 1,765 protein sequences across multiple bacterial species, categorized into two classes (380 T4SEs as the positive class and 1,385 non-T4SEs as the negative class). These proteins in this dataset have mutual sequence identity no more than 30%. The 1,765 protein sequences were divided into two subsets for cross validation in the training and the independent testing, respectively. The training dataset (Train-915) are composed of 915 sequences, among which 305 T4SE sequences were randomly selected from positive class, and 610 non-T4SE sequences were randomly selected from negative class. The dataset of Train-915 was further randomly divided into five subsets (or folds) with an equal number of protein sequences for cross validation to attain the optimized model. In each of the five validations, 4 of the 5-folds were used for training and the remaining one for testing, which was repeated for five times. The testing dataset (Test-850) included the remaining 75 T4SE sequences as positive samples and 775 non-T4SE sequences as negative samples for independent testing.

Feature Representation of Protein Sequence Samples

One of the key problems in designing a predictor based on machine learning is how to encode a protein sequence as an informative feature vector enriched with highly discriminative information. In the present section, we describe how to formulate an effective mathematical expression that describes protein sequences in the training and testing data sets.

The protein sequence profile (i.e., PSSM) is a powerful representation of residue or sequence information of proteins. It has achieved good performance on a number of bioinformatics applications such as functional residues prediction and protein function prediction (Xiong et al., 2011a,b, 2012; Zhu et al., 2013; Wei et al., 2017a). In this study, PSSMs were generated by three iterations of PSI-BLAST searches against Uniref50 with the BLOSUM62 substitution matrix. The parameter of e-value was set to 0.001. Because ML-based models can only handle vectors with equal lengths for all protein sequence samples, the PSSM of a protein sequence (amino acid length is L) has a dimension of L*20, which could not be directly used as the input feature vector for machine learning algorithms. Instead, the original PSSM profile was further used to calculate the feature of PSSM-composition by summing rows that correspond to the same amino acid residues in a PSSM profile, in much the same way as the previous studies (Zou et al., 2013; Wang J. et al's., 2017). The sum value was divided by the length of the protein sequence for each type of amino acid (there is a total of 20 types). Thus, a vector of size at 400 (=20 × 20) is finally used for representing a protein sequence sample. **Figure 1** presents the details about how to generate a feature vector of PSSM-composition for a given protein sequence.

Classification System

The ensemble learning techniques can be categorized into three main types, which include bagging, boosting, and stacked ensemble. It is demonstrated that the ensemble learning techniques can help improve the prediction performance in various bioinformatics applications (Zhang et al., 2012; Lin et al., 2013, 2014; Zou et al., 2015; Li et al., 2016; Yuan et al., 2016; Wan et al., 2017; Iqbal and Hoque, 2018; Mishra et al., 2018; You et al., 2018). In this section, we introduce the components of the two-stage stacked ensemble scheme, including various classification algorithms used as base-classifiers in the first stage, and the input of the meta-classifier in the second stage.

Base-Classifier

In order to find the optimal combination of base-classifiers in the first stage and the meta-classifier in the second stage, the following eight different machine learning algorithms were exploited: (i) SVM (Cortes and Vapnik, 1995), (ii) Naïve Bayes (NB), (iii) K Nearest Neighbor (KNN), (iv) Logistic Regression (LR), (v) Random Forest (RF) (Breiman, 2001), (vi) Extremely Randomized Trees (ERT) (Geurts et al., 2006), (vii) Gradient Boosting Machine (GBM) (Friedman, 2001), and (viii) eXtreme Gradient Boosting (XGB). The algorithms such as NB, LR, and GBM were implemented by using h2o package in R software. The algorithms of SVM, KNN, RF, ERT, and XGB are implemented by using e1071, caret, randomForest, extraTrees and xgboost packages in R, respectively. The optimal parameters in these algorithms are determined by a grid search strategy.

Meta-Classifier

The meta-classifier in the second level generalization (or stacked generalization) is used to combine the outputs of base-classifiers in an ensemble. In our classification system, we applied a stacked generalization approach proposed by Wolpert (1992), in which an ensemble of base-classifiers are first constructed, whose outputs are used as inputs to a second level of meta-classifier to learn the relationship between the ensemble outputs and the actual classes/labels. The stacked generalization scheme can be viewed as an extension version of cross validation. In the first stage, the base-classifiers were trained with the feature of PSSM-composition of sequences. In the second stage, the prediction class probabilities of the base-classifiers were taken as inputs to the meta-classifier (shown in **Figure 2**).

Model Validation Method

To evaluate performances of classification models, the validation methods are mainly consisting of k-fold cross validation, leave-one-out cross validation (or called as jackknife test), and independent tests. In k-fold cross validation, the sample set is randomly divided into k subsets with equal sizes. Of the k subsets, only one subset is selected as the validation data for testing the model, and the remaining k-1 subsets are used as training data. The cross validation process is then repeated k times (the folds), with each of the k subsets used exactly once as the validation data. The results from k folds are finally averaged. The k-fold cross validation method has been widely used as the model validation approach in various bioinformatics applications (Zhu and Mitchell, 2011; Xu et al., 2017; Zeng et al., 2017; Chen X. et al., 2018; He et al., 2018a,d). In the present study, the 5-fold cross validation was used for validation in the training set, and the independent test was used for testing the generalization ability of the proposed method, and comparison with other methods.

Model Evaluation Metric

In order to assess prediction performances of single-label classification systems, a set of six threshold-dependent metrics are widely used in the bioinformatics studies (Xia et al., 2010; Li et al., 2011; Zhang et al., 2017, 2018a,b,c; He et al., 2018c; Jia et al., 2018; Zhao et al., 2018). They are accuracy (ACC), sensitivity (SE, also called recall), specificity (SP), precision (PR), Matthew's correlation coefficient (MCC) and F-measure (F_1). The definitions of these metrics are shown as below.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$SE = \frac{TP}{TP + FN} \tag{2}$$

$$SP = \frac{TN}{TN + FP} \tag{3}$$

$$PR = \frac{TP}{TP + FP} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{5}$$

$$F_1 = \frac{2 \times SE \times PR}{SE + PR} \tag{6}$$

where TP (true positives) is the number of correctly predicted T4SEs, TN (true negatives) is the number of correctly predicted

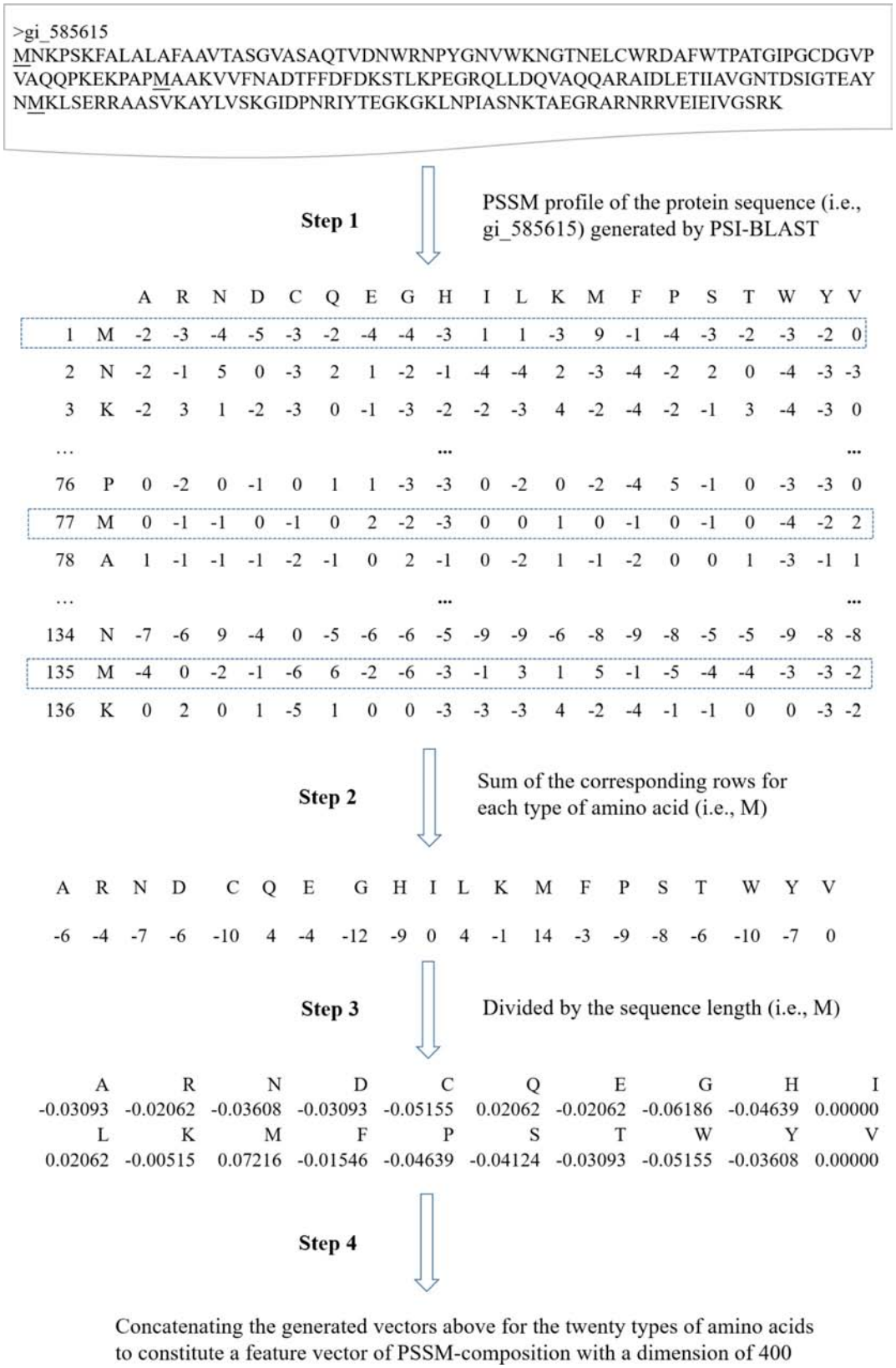


FIGURE 1 | The illustration of PSSM-composition profile calculation for a query sequence.

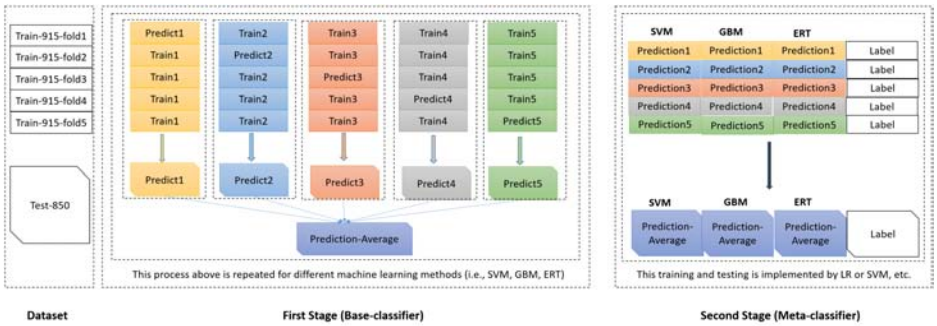


FIGURE 2 | The framework of the stacked ensemble scheme proposed in PredT4SE-Stack.

TABLE 1 | Performance comparison of eight types of base-classifiers in the first stage on Train-915 dataset using 5-fold cross validation.

Method	Parameter	ACC(%)	SE (%)	SP (%)	PR(%)	MCC	F_1
NB	laplace = 0	73.2	81.0	69.3	57.0	0.476	0.669
KNN	k = 10	85.5	82.0	87.2	76.3	0.680	0.790
LR	family = "binomial"	87.9	74.8	94.4	87.1	0.722	0.803
RF	ntree = 500	88.5	72.5	96.6	91.4	0.738	0.807
ERT	numRandomCuts = 9	89.4	74.8	96.7	92.1	0.759	0.824
SVM	cost = 1, gamma = 2^{-8}, kernel = "radial"	90.2	78.0	96.2	91.6	0.777	0.839
XGB	eta = 0.3, max_depth = 6, nrounds = 500, objective = "binary:logistic"	90.1	78.7	95.7	90.4	0.774	0.840
GBM	learn_rate = 0.7, max_depth = 9, ntrees = 50	90.5	80.0	95.7	90.7	0.784	0.847

non-T4SEs, FP (false positives) is the number of non-T4SEs wrongly predicted as T4SEs, and FN (false negatives) is the number of T4SEs wrongly predicted as non-T4SEs.

The receiver operating characteristic (ROC) curve is a plot of the sensitivity versus (1-specificity) for a binary classifier at varying thresholds from 0 to 1 (the threshold is assigned as the probability of the target sequence to be a T4SE in our study). The area under the curve (AUC) can be used as a powerful metric for evaluation performances of classifiers. It is worth mentioning that AUC of ROC (and ACC, MCC) can present overly optimistic assessment of performance of an algorithm on a heavily unbalanced dataset. Therefore, we only used AUC of ROC for evaluation in 5-fold cross validation, but not used it for evaluation in the independent dataset (only 75 proteins are true positives among 850 samples). Instead, the metric of F_1, which is a harmonic mean of recall (or sensitivity) and precision, is a main metric for evaluating performances of classifiers in the present study.

RESULTS AND DISCUSSION

Predictive Power of Various Base-Classifiers on Train-915 Dataset

The aim of this section is to test the predictive power of base-classifiers based on PSSM-composition profiles for eight different machine learning algorithms on Train-915 dataset using 5-fold cross validation. Experimental results shown in **Table 1** indicate that the algorithm of naïve Bayes performed worst on this task. The algorithms of KNN, logistic regression, random forest, and extremely randomized trees performed moderately. The algorithms of support vector machine, extreme gradient boosting, and gradient boosting machine performed best. The results of ROC shown in **Figure 3** are mainly in agreement with the findings in **Table 1**. However, the fact that the AUC-ROC of SVM is higher than that of XGB and GBM indicates that SVM can achieve more stable performance than XGB and GBM using PSSM-composition feature as input in the present task, in regardless of the change of the thresholds. It should be noted that we tried a large number of other types of PSSM-derived features generated by POSSUM tookit (Wang J. et al's., 2017), and a variety of structural and physiochemical descriptors extracted from protein sequences generated by iFeature tookit (Chen Z. et al., 2018) when we designed the input features of the base-classifiers. Our experimental results demonstrated that the PSSM-composition feature utilized in this study yielded satisfactory performance, which performed better than other types of sequence-based features. Moreover, we attempted to directly combine the PSSM-composition feature with other types of features as the input of the base-classifiers. It was found that the combined features could not significantly produce higher performance than the single type of PSSM-composition feature (data not shown).

Predictive Power of Meta-Classifiers on Train-915 Dataset

Since combining all of the above mentioned base-classifiers in a meta-classifier could not yield optimal prediction performance, it is desirable to search for the optimal combination of base-classifiers. Since RF and ERT are tree-based classifiers, we chose

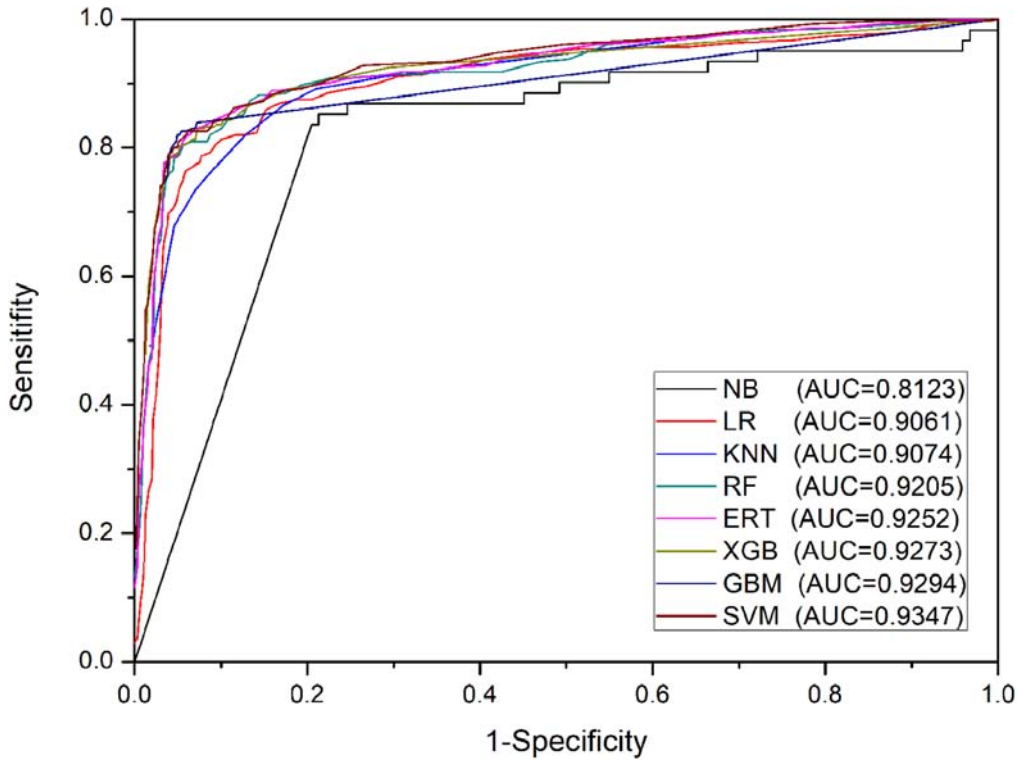


FIGURE 3 | ROC curves of base-classifiers in the first stage on Train-915 dataset using 5-fold cross validation.

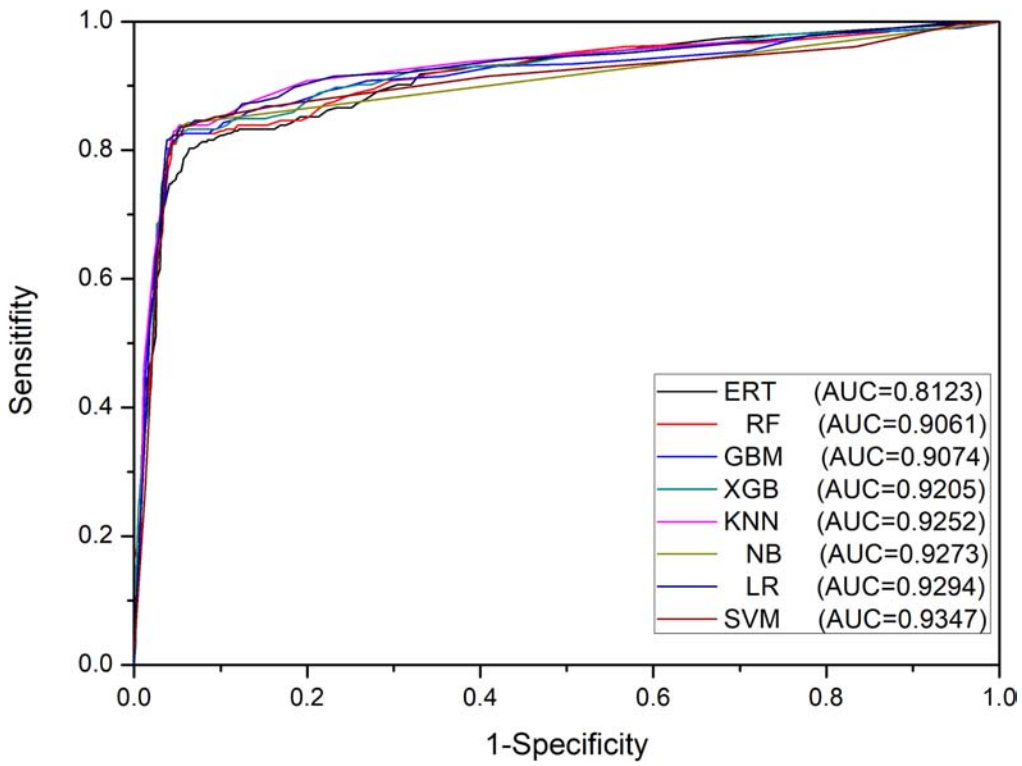


FIGURE 4 | ROC curves of meta-classifiers in the second stage on Train-915 dataset using 5-fold cross validation.

TABLE 2 | Performance comparison of eight types of meta-classifiers in the second stage on Train-915 dataset using 5-fold cross validation.

Method	Parameter	ACC(%)	SE (%)	SP (%)	PR(%)	MCC	F_1
ERT	numRandomCuts = 9	88.9	80.3	93.1	86.5	0.752	0.828
RF	ntree = 500	90.4	81.0	95.1	89.7	0.783	0.847
SVM	cost = 10, gamma = 2^{-10}, kernel = "radial"	90.6	80.3	95.7	90.7	0.787	0.849
GBM	learn_rate = 0.1, max_depth = 3, ntrees = 50	90.6	82.0	94.9	89.3	0.788	0.851
XGB	eta = 0.1, max_depth = 2, nrounds = 100, objective = "binary:logistic"	90.7	81.3	95.4	90.4	0.791	0.852
NB	laplace = 0	90.9	82.3	95.2	89.9	0.795	0.857
KNN	k = 19	91.0	82.0	95.6	90.5	0.797	0.857
LR	family = "binomial"	91.1	81.0	96.2	91.9	0.800	0.858

one of them at a time. Because GBM and XGB are boosting-based methods, and XGB is an efficient and scalable implementation of GBM, we chose one of them too. It was found that the combination of SVM, GBM, and ERT achieved the optimal performance, which is in agreement with the finding of study by Pan et al. (2018) on the prediction task of hot spots in protein-RNA interfaces.

Furthermore, we tested the same set of eight ML methods as the classification algorithms of meta-classifiers to compare their prediction performances. The results in **Table 2** showed that all meta-classifiers except the one based on ERT achieved very similar performances, for example, the values of F_1 are falling in a narrow range from 0.847 to 0.858, whereas the base-classifiers using the same set of ML algorithms are ranging from 0.669 to 0.847 in the first stage. These results can be explained by the fact that the pattern learned from the first stage is effective enough, leading to the similar level of performances at the second stage on the same dataset of Train-915, irrespective of ML algorithms, except ERT (also demonstrated in **Figure 4**) .

Predictive Power of Meta-Classifiers on Test-850 Dataset

In the section, the prediction performances of meta-classifiers are evaluated on the independent dataset, which is mimicking a true prediction task, since the model trained on one dataset is really tested on an unseen dataset for examining its generalization ability on a new dataset. **Table 3** indicated that LR and SVM have top performances on Test-850 dataset. Therefore, both of them can be utilized as the classification algorithms of the meta-classifier in PredT4SE-Stack. Considering the fact that LR is more interpretable than SVM, we could use LR to construct the meta-classifier in our model PredT4SE-Stack. In real application, we will re-train PredT4SE-Stack on a whole dataset consisting of Train-915 and Test-850.

Comparison With Previous Studies

The main purpose of this section is to compare our proposed approach PredT4SE-Stack to previously published methods. Performance comparisons among different T4SE prediction approaches are scientifically meaningful only if they train and test their methods on the same dataset. Accordingly, our approach PredT4SE-Stack was only compared with the recently published method proposed by Wang Y. et al. (2017). The first reason is that both two studies used the same benchmark dataset for training and testing. The second reason is that Wang Y. et al.'s (2017) method had been proved to be improved over other published methods such as T4EffPred (Zou et al., 2013), T4SEpre (Wang et al., 2014), and An et al.'s (2016) method. **Table 4** shows the comparison results between our method with Wang Y. et al.'s (2017) method. Since the measures of F_1 and precision are not available in **Table 4** in their published study, we firstly calculated the TP, TN, FP, and FN using the sensitivity and specificity of their method, and then calculated F_1 and precision of Wang Y. et al.'s (2017) method. The meta-classifier of our PredT4SE-Stack classification system was implemented by SVM or LR. For SVM or LR, the performance (F_1 = 0.734 or 0.733) of our method is much higher than that (F_1 = 0.521) of Wang Y. et al.'s (2017) method. If our SVM-based meta-classifier is tuned on the same recall or sensitivity of 90.7%, our method achieved better performance at specificity, precision,

TABLE 3 | Performance comparison of eight types of meta-classifiers in the second stage on the independent dataset Test-850.

Method	ACC(%)	SE (%)	SP (%)	PR(%)	MCC	F_1
XGB	92.4	85.3	93.0	54.2	0.643	0.663
GBM	93.1	88.0	93.5	56.9	0.674	0.691
KNN	93.5	88.0	94.1	58.9	0.688	0.706
RF	93.8	86.7	94.5	60.2	0.691	0.710
NB	93.8	88.0	94.3	60.0	0.696	0.714
ERT	94.0	88.0	94.6	61.1	0.703	0.721
LR	94.4	88.0	95.0	62.9	0.715	0.733
SVM	94.5	86.7	95.2	63.7	0.715	0.734

TABLE 4 | Performance comparison between our method with the other method on the independent dataset Test-850.

Method	ACC(%)	SE (%)	SP (%)	PR(%)	MCC	F_1
Wang Y. et al.'s (2017) method	85.3	90.7	84.8	36.6	0.518	0.521
PredT4SE-Stack (SVM, 0.23)	87.5	90.7	87.2	40.7	0.556	0.562
PredT4SE-Stack (SVM, 0.50)	94.5	86.7	95.2	63.7	0.715	0.734
PredT4SE-Stack (LR, 0.11)	88.7	90.7	88.5	43.3	0.579	0.586
PredT4SE-Stack (LR, 0.50)	94.4	88.0	95.0	62.9	0.715	0.733

and F_1, which are 2.4, 4.1, and 4.1% respectively, higher than that of Wang Y. et al.'s (2017) method. If our LR-based meta-classifier is tuned on the same recall or sensitivity of 90.7%, our method achieved better performance at specificity, precision, and F_1, which are 3.7, 6.7, and 6.5% respectively, higher than that of Wang Y. et al.'s (2017) method.

CONCLUSION

The main goal of the current study is to develop a stacked ensemble model PredT4SE-Stack to predict T4SEs from protein sequence information. The proposed model utilized an ensemble of base-classifiers implemented by SVM, GBM, and ERT to generate outputs for the meta-classifier in the classification system. It was demonstrated that the framework of PredT4SE-Stack was a feasible and effective way to accurately identify T4SEs based on protein sequence information. However, the performance of PredT4SE-Stack can be further improved in several respects. Firstly, the diversity of base-classifiers was implemented by various classification algorithms in the present work. It can be further improved by different features in different base-classifiers. Secondly, inspired by the successful application of feature selection strategies in various bioinformatics tasks (Zou et al., 2016; Wei et al., 2017b, 2018; He et al., 2018b; Manavalan et al., 2018; Qiao et al., 2018; Su et al., 2018; Tang et al., 2018), the predictive power of base-classifiers can be boosted by incorporating an effective feature selection technology on a large pool of sequence-derived features. Moreover, an effective model selection on a large number of candidate base-classifiers will be explored to improve the prediction performance of the meta-classifier. These improvements will be explored in the further study.

AUTHOR CONTRIBUTIONS

XZ and D-QW conceived the study. YX and XZ designed the experiments. YX performed the experiments. YX, QW, JY, and XZ analyzed the data. YX and XZ wrote paper. All authors reviewed the manuscript and agreed to this information prior to submission.

FUNDING

This work was supported by the funding from National Natural Science Foundation of China for Young Scholars (Grant Nos. 31601074 and 21403002), National Key Research Program (Contract No. 2016YFA0501703), Shanghai Jiao Tong University School of Medicine (Contract Nos. YG2015QN34 and YG2017ZD14), and Shanghai Key Laboratory of Intelligent Information Processing (Contract No. IIPL-2016-005).

REFERENCES

An, Y., Wang, J., Li, C., Leier, A., Marquez-Lago, T., Wilksch, J., et al. (2016). Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI. *Brief Bioinform.* 19, 148–161. doi: 10.1093/bib/bbw100

An, Y., Wang, J., Li, C., Revote, J., Zhang, Y., Naderer, T., et al. (2017). SecretEPDB: a comprehensive web-based resource for secreted effector proteins of the bacterial types III, IV and VI secretion systems. *Sci. Rep.* 7:41031. doi: 10.1038/srep41031

Bi, D., Liu, L., Tai, C., Deng, Z., Rajakumar, K., and Ou, H. Y. (2013). SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic Acids Res.* 41, D660–D665. doi: 10.1093/nar/gks1248

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Burstein, D., Zusman, T., Degtyar, E., Viner, R., Segal, G., and Pupko, T. (2009). Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog.* 5:e1000508. doi: 10.1371/journal.ppat.1000508

Chandran, V., Fronzes, R., Duquerroy, S., Cronin, N., Navaza, J., and Waksman, G. (2009). Structure of the outer membrane complex of a type IV secretion system. *Nature* 462, 1011–1015. doi: 10.1038/nature08588

Chen, C., Banga, S., Mertens, K., Weber, M. M., Gorbaslieva, I., Tan, Y., et al. (2010). Large-scale identification and translocation of type IV secretion substrates by *Coxiella burnetii*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21755–21760. doi: 10.1073/pnas.1010485107

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018). BNPMDA: bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Eichinger, V., Nussbaumer, T., Platzer, A., Jehl, M. A., Arnold, R., and Rattei, T. (2016). EffectiveDB–updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res.* 44, D669–D674. doi: 10.1093/nar/gkv1269

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Fronzes, R., Schafer, E., Wang, L., Saibil, H. R., Orlova, E. V., and Waksman, G. (2009). Structure of a type IV secretion system core complex. *Science* 323, 266–268. doi: 10.1126/science.1166101

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

He, B. S., Qu, J., and Zhao, Q. (2018a). Identifying and exploiting potential miRNA-disease associations with neighborhood regularized logistic matrix factorization. *Front. Genet.* 9:303. doi: 10.3389/fgene.2018.00303

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018b). PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics* 19:306. doi: 10.1186/s12859-018-2321-0

He, W., Jia, C., and Zou, Q. (2018c). 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* doi: 10.1093/bioinformatics/bty668 [Epub ahead of print].

He, W., Ju, Y., Zeng, X. X., Liu, X. R., and Zou, Q. (2018d). ScncDNAPred: a sequence-based predictor for identifying non-coding DNA in *Saccharomyces cerevisiae*. *Front. Microbiol.* 9:2174. doi: 10.3389/fmicb.2018.02174

Iqbal, S., and Hoque, M. T. (2018). PBRpredict-Suite: a suite of models to predict peptide-recognition domain residues from protein sequence. *Bioinformatics* 34, 3289–3299. doi: 10.1093/bioinformatics/bty352

Jia, C., Zuo, Y., and Zou, Q. (2018). O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics* 34, 2029–2036. doi: 10.1093/bioinformatics/bty039

Li, D., Luo, L., Zhang, W., Liu, F., and Luo, F. (2016). A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinformatics* 17:329. doi: 10.1186/s12859-016-1206-3

Li, J., Yao, Y., Xu, H. H., Hao, L., Deng, Z., Rajakumar, K., et al. (2015). SecReT6: a web-based resource for type VI secretion systems found in bacteria. *Environ. Microbiol.* 17, 2196–2202. doi: 10.1111/1462-2920.12794

Li, Z., Wong, L., and Li, J. (2011). DBAC: a simple prediction method for protein binding hot spots based on burial levels and deeply buried atomic contacts. *BMC Syst. Biol.* 5(Suppl. 1):S5. doi: 10.1186/1752-0509-5-S1-S5

Lifshitz, Z., Burstein, D., Peeri, M., Zusman, T., Schwartz, K., Shuman, H. A., et al. (2013). Computational modeling and experimental validation of the *Legionella* and *Coxiella* virulence-related type-IVB secretion signal. *Proc. Natl. Acad. Sci. U.S.A.* 110, E707–E715. doi: 10.1073/pnas.1215278110

Lin, C., Chen, W. Q., Qiu, C., Wu, Y. F., Krishnan, S., and Zou, Q. (2014). LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing* 123, 424–435. doi: 10.1016/j.neucom.2013.08.004

Lin, C., Zou, Y., Qin, J., Liu, X., Jiang, Y., Ke, C., et al. (2013). Hierarchical classification of protein folds using a novel ensemble classifier. *PLoS One* 8:e56499. doi: 10.1371/journal.pone.0056499

Manavalan, B., Shin, T. H., and Lee, G. (2018). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Mishra, A., Pokhrel, P., and Hoque, M. T. (2018). StackDPPred: a stacking based prediction of dna-binding protein from sequence. *Bioinformatics* doi: 10.1093/bioinformatics/bty653 [Epub ahead of print].

Pan, Y., Wang, Z., Zhan, W., and Deng, L. (2018). Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 34, 1473–1480. doi: 10.1093/bioinformatics/btx822

Qiao, Y., Xiong, Y., Gao, H., Zhu, X., and Chen, P. (2018). Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. *BMC Bioinformatics* 19:14. doi: 10.1186/s12859-018-2009-5

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* doi: 10.1093/bioinformatics/bty508 [Epub ahead of print].

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17:1700262. doi: 10.1002/pmic.201700262

Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T. T., Webb, G., et al. (2017). POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics* 33, 2756–2758. doi: 10.1093/bioinformatics/btx302

Wang, Y., Guo, Y., Pu, X., and Li, M. (2017). Effective prediction of bacterial type IV secreted effectors by combined features of both C-termini and N-termini. *J. Comput. Aided Mol. Des.* 31, 1029–1038. doi: 10.1007/s10822-017-0080-z

Wang, Y., Wei, X., Bao, H., and Liu, S. L. (2014). Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics* 15:50. doi: 10.1186/1471-2164-15-50

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2018). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* doi: 10.1093/bioinformatics/bty824 [Epub ahead of print].

Wei, L., Tang, J. J., and Zou, Q. (2017a). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026

Wei, L., Xing, P., Shi, G., Ji, Z. L., and Zou, Q. (2017b). Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2017.2670558 [Epub ahead of print].

Wolpert, D. H. (1992). Stacked generalization. *Neural Netw.* 5, 241–259. doi: 10.1016/S0893-6080(05)80023-1

Xia, J. F., Zhao, X. M., Song, J., and Huang, D. S. (2010). APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics* 11:174. doi: 10.1186/1471-2105-11-174

Xiong, Y., Liu, J., and Wei, D. Q. (2011a). An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins* 79, 509–517. doi: 10.1002/prot.22898

Xiong, Y., Xia, J., Zhang, W., and Liu, J. (2011b). Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures. *PLoS One* 6:e28440. doi: 10.1371/journal.pone.0028440

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Sci.* 10(Suppl. 1):S20. doi: 10.1186/1477-5956-10-S1-S20

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., et al. (2017). PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.* 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., and Zhu, S. (2018). GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* 34, 2465–2473. doi: 10.1093/bioinformatics/bty130

Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S. (2016). DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 32, i18–i27. doi: 10.1093/bioinformatics/btw244

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 687–695. doi: 10.1109/TCBB.2016.2520947

Zhang, J., Feng, P., Lin, H., and Chen, W. (2018a). Identifying RNA N(6)-Methyladenosine Sites in *Escherichia coli* genome. *Front. Microbiol.* 9:955. doi: 10.3389/fmicb.2018.00955

Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., and Li, X. (2018b). Feature-derived graph regularized matrix factorization for predicting drug side effects. *Neurocomputing* 287, 154–162. doi: 10.1016/j.neucom.2018.01.085

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018c). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhang, W., Chen, Y., Liu, F., Luo, F., Tian, G., and Li, X. (2017). Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinformatics* 18:18. doi: 10.1186/s12859-016-1415-9

Zhang, W., Niu, Y., Xiong, Y., Zhao, M., Yu, R., and Liu, J. (2012). Computational prediction of conformational B-cell epitopes from antigen primary structures by ensemble learning. *PLoS One* 7:e43575. doi: 10.1371/journal.pone.0043575

Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018). IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zhu, X., Ericksen, S. S., and Mitchell, J. C. (2013). DBSI: DNA-binding site identifier. *Nucleic Acids Res.* 41:e160. doi: 10.1093/nar/gkt617

Zhu, X., and Mitchell, J. C. (2011). KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins* 79, 2671–2683. doi: 10.1002/prot.23094

Zou, L., Nan, C., and Hu, F. (2013). Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* 29, 3135–3142. doi: 10.1093/bioinformatics/btt554

Zou, Q., Guo, J., Ju, Y., Wu, M., Zeng, X., and Hong, Z. (2015). Improving tRNAscan-SE annotation results via ensemble classifiers. *Mol. Inform.* 34, 761–770. doi: 10.1002/minf.201500031

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10(Suppl. 4):114. doi: 10.1186/s12918-016-0353-5

Conflict of Interest Statement: 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.

Copyright © 2018 Xiong, Wang, Yang, Zhu and Wei. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 29 October 2018
doi: 10.3389/fmicb.2018.02320

A Cross-Scale Neutral Theory Approach to the Influence of Obesity on Community Assembly of Human Gut Microbiome

Wendy Li[1], *Yali Yuan*[1,2], *Yao Xia*[1,3], *Yang Sun*[4], *Yinglei Miao*[4] *and Sam Ma*[1,5]*

[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] *College of Clinical Medicine, Lanzhou University, Lanzhou, China,* [3] *Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming, China,* [4] *Department of Gastroenterology, The First Affiliated Hospital of Kunming Medical University, Yunnan Institute of Digestive Disease, Kunming, China,* [5] *Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China*

Background: The implications of gut microbiome to obesity have been extensively investigated in recent years although the exact mechanism is still unclear. The question whether or not obesity influences gut microbiome assembly has not been addressed. The question is significant because it is fundamental for investigating the diversity maintenance and stability of gut microbiome, and the latter should hold a key for understanding the etiological implications of gut microbiome to obesity.

Methods: In this study, we adopt a dual neutral theory modeling strategy to address this question from both species and community perspectives, with both discrete and continuous neutral theory models. The first neutral theory model we apply is Hubbell's neutral theory of biodiversity that has been extensively tested in macro-ecology of plants and animals, and the second we apply is Sloan's neutral theory model that was developed particularly for microbial communities based on metagenomic sequencing data. Both the neutral models are complementary to each other and integrated together offering a comprehensive approach to more accurately revealing the possible influence of obesity on gut microbiome assembly. This is not only because the focus of both neutral theory models is different (community vs. species), but also because they adopted two different modeling strategies (discrete vs. continuous).

Results: We test both the neutral theory models with datasets from Turnbaugh et al. (2009). Our tests showed that the species abundance distributions of more than ½ species (59–69%) in gut microbiome satisfied the prediction of Sloan's neutral theory, although at the community level, the number of communities satisfied the Hubbell's neutral theory was negligible (2 out of 278).

Conclusion: The apparently contradictory findings above suggest that both stochastic neutral effects and deterministic environmental (host) factors play important roles in shaping the assembly and diversity of gut microbiome. Furthermore, obesity may just

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Jianghan Qu,
University of Southern California, United States
Jing Lu,
Walmart Labs, United States

***Correspondence:**
Sam Ma
samma@uidaho.edu

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

Received: *31 May 2018*
Accepted: *11 September 2018*
Published: *29 October 2018*

Citation:
Li W, Yuan Y, Xia Y, Sun Y, Miao Y and Ma S (2018) A Cross-Scale Neutral Theory Approach to the Influence of Obesity on Community Assembly of Human Gut Microbiome. Front. Microbiol. 9:2320. doi: 10.3389/fmicb.2018.02320

be one of the host factors, but its influence may not be strong enough to tip the balance between stochastic and deterministic forces that shape the community assembly. Finally, the apparent contradiction from both the neutral theories should not be surprising given that there are still near 30–40% species that do not obey the neutral law.

Keywords: obesity, Hubbell neutral theory of biodiversity, Sloan's neutral model for microbes, niche theory, community assembly, species abundance distribution (SAD)

INTRODUCTION

Obesity is a complex physiological disorder that is often associated with multi-organ (e.g., cardiac, adipose, muscle, hypothalamic, pancreatic, and hepatic tissue), chronic metabolic, and inflammatory alterations. Obesity may induce some chronic metabolic diseases directly or indirectly, such as type 2 diabetes, atherosclerosis, nonalcoholic fatty liver disease, and gout (Sun et al., 2012; Henao-Mejia et al., 2014). Obesity has become a serious health threat to a growing number of people around the world in the past decades. Obesity epidemic relates to many factors, including not only diet habits, physical activity, and genetic makeup (Ravussin and Ryan, 2018), but also behavioral factors, environmental exposures, social-psychological factors, and reproductive factors (Davis et al., 2018). In addition, its close links with the human gut microbiome have been revealed by more recent studies in the last decade (e.g., Turnbaugh et al., 2006, 2009; Zhao, 2013; Davis et al., 2018). Because of the significant overlap between obesity and the metabolic syndrome, dysbiosis of gut microbiome or shift of the balance, is a phenomenon deserving serious considerations when assessing the elements driving adiposity (Stephens et al., 2018). Several studies showed a significant difference in the ratio of *Firmicutes* to *Bacteroidetes*, where higher *Firmicutes* and lower *Bacteroidetes* were found in obese subjects (Ley et al., 2005, 2006; Turnbaugh et al., 2006, 2009; Armougom et al., 2009; Hildebrandt et al., 2009; Fleissner et al., 2010; Murphy et al., 2010), but exceptions regarding the ratio change were also reported (Schwiertz et al., 2010; Zhang et al., 2010; Zhao, 2013). More recent studies found that the abundance of *Bacteroides thetaiotaomicron* remarkably decreased in obese individuals (Liu et al., 2017), and the ratio of two enterotypes in human gut microbiome (*Prevotella* spp. to *Bacteroides* spp.) has been shown to play a role in predicting the weight loss of people with different diets (Hjorth et al., 2017). Goodrich et al. found that the family *Christensenellaceae* was enriched in individuals with low body mass index (BMI), and the weight is reduced in the recipient mice inoculated with *Christensenella minuta* (Goodrich et al., 2014, 2016a,b). Menni et al. (2017) further assessed the association of gut microbiome composition and change in body weight over time by analyzing the data of 1632 females from "TwinsUK" database including longitudinal BMI data and fecal microbiome data. They demonstrated that *Ruminococcaceae* and *Lachnospiraceae* were associated with lower long-term weight gain, and *Bacterioides* was associated with increased risk of weight gain. In addition, many studies have suggested the lowered gut microbial diversity in obese individuals (Ley et al., 2006; Turnbaugh et al., 2009; Le Chatelier et al., 2013). In spite of the extensive studies on the relationship between gut microbiome diversity and obesity, and several computational models that can help for predicting potential obesity-related microbe (Chen et al., 2017; Huang et al., 2017a,b; Wang et al., 2017), the underlying mechanism has not been addressed to the best of our knowledge.

The mechanisms of species coexistence and biodiversity maintenance in ecological communities have long been a core research theme of community ecology, in which the deterministic niche theory and stochastic neutral theory are well recognized as two most influential. Traditional niche theory maintains that species coexisting in a community must have different niches, and species with the same niche requirements could not stably coexist in long term (Matthews et al., 2014). Although niche theory was supported by many field and laboratory studies, it encountered difficulties in explaining the mechanisms of species coexistence in tropical forests. Hubbell (1997, 2001); Wills et al. (1997) introduced the neutral theory of biodiversity that provided alternative perspectives of species coexistence. Hubbell's neutral theory of biodiversity is an individual-based stochastic dynamic theory that assumes equivalences among interacting species and can be formulated as a dispersal-limited, distribution-sampling model (Etienne, 2005; Alonso et al., 2006; Rosindell et al., 2011, 2012). The latter allows rigorous statistical testing of the neutral theory with the species abundance data (SAD) that can be obtained from field survey (in macro-ecology of plants and animals) or metagenomic sequencing data (in microbial ecology).

In consideration of the unique characteristics of metagenomic sequencing data of microbial species abundance distribution, Sloan et al. (2006, 2007) proposed an alternative neutral model that emphasizes the species-level neutrality in microbial communities. Unlike traditional neutral theories that were calibrated by using "almost complete description of the taxa-abundance distribution for community," Sloan's model can calibrate itself just with the small-sample microbial data that were collected using molecular approaches since Sloan's model allowed for the difference of competitiveness among species in microbial communities (Sloan et al., 2006, 2007). Another important characteristic of Sloan's model is that it was derived from a continuous diffusion process rather than from a discrete distribution model as that of Hubbell (Sloan et al., 2006, 2007). These two features make Sloan's neutral model a nice complement to Hubbell's neutral model (Hubbell, 2001; Etienne, 2005; Rosindell et al., 2011, 2012).

The neutral theory offers a powerful quantitative tool to identify the forces that shape the gut microbial communities, and the revealed information is crucial for understanding

the mechanisms that maintain microbiome diversity and possible influences of diseases/disorders such as obesity on the mechanistic shifts of community assembly. In spite of extensive studies on the relationship between the gut microbiome and obesity, as reviewed previously, whether or not obesity plays a tipping role in "re-assembling" gut microbiota, or exerting a significant influence on the mechanisms of community assembly and diversity maintenance, is still an open question. For example, the test of neutral theory can help to answer the following question: which forces, deterministic host factors such as obesity, or stochasticities in birth, death and migration of gut microbes, are in control of the composition and diversity of gut microbiome. If the former is the case, it suggests that the community is formed through the partition of different niches, occupied by species with different niche requirements, and the exhibited diversity (heterogeneity) at the community level is determined by the deterministic environmental factors that delineate different niches. If the latter is the case, it suggests that the community is essentially a random mix of largely ecological equivalent species, and the exhibited diversity (heterogeneity) is caused by the stochasticities in birth, death and migration of different species. The primary objective of this article is to apply the neutral theories of Hubbell (2001) and Sloan et al. (2006, 2007) for exploring the above question with the dataset from a large-scale, comprehensive study of the human gut microbiome involving 283 overweight, obese and lean individuals, originally reported by Turnbaugh et al. (2009).

MATERIAL AND METHODS

Dataset Description

The 16S r-RNA datasets of gut microbiomes we used to test the neutral theories were first reported in Turnbaugh et al. (2009), and a brief description is presented as follows. A series of fecal samples were collected from 154 individuals, including 31 monozygotic twin pairs, 23 dizygotic twin pairs and their mothers ($n = 46$), and each participant was sampled twice with an average interval between sampling of 57 ± 4 days. A total of 283 fecal samples were taken, including 196 were collected from participants in obesity (BMI $>$ 30 kg/m^{-2}), 61 were collected from participants in leanness, and 24 were collected from overweight participants (BMI $\geq$ 25 and $<$ 30). The datasets of 16S rRNA reads and corresponding species or OTU (operational taxonomic unit) table was obtained by using the 454 FLX platform and subsequent bioinformatics analysis. Each sample corresponds to one row in the OTU table, and was treated as one microbial community. More detailed information on the dataset is referred to Turnbaugh et al. (2009).

Hubbell's (2001) Neutral Theory Model

Hubbell's neutral theory is an individual-based sampling theory, and offers a biological occurrence mechanism to explain observed species abundance distributions (SADs) in ecological communities. It assumes that all individuals in a saturated local community are ecologically equivalent, which means they have the same rate of birth, death and migration, excluding their random fluctuations. Etienne (2005) developed a sampling formula (distribution) that can be utilized to statistically test the Hubbell' neutral theory with field observation data of SAD, in our case the OTU tables described in the previous section.

Etienne sampling formula (Etienne, 2005) is with the following form:

$$P(D|\theta, m, J) = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J} \phi_j!} \frac{\theta^S}{(I)_J} \sum_{A=S}^{J} K(D, A) \frac{I^A}{(\theta)_A}, \quad (1)$$

where m is the migration probability, J is the total number of individuals in the community, I is the number of immigrants that compete with the local community individuals, S is the total number of species, θ is the fundamental biodiversity parameter of the formula, n_i is the abundance of species i, ϕ_j is the number of species with abundance j, D is the species-abundance distribution containing the abundance of each species, $D = (n_1, n_2, ..., n_s)$.

The immigration rate (probability) m is further defined as:

$$m = \frac{I}{I + j - 1}, \quad (2)$$

$K(D, A)$ is further defined as:

$$K(D, A) = \sum_{\{a_1, a_2 .... a_S | \sum_{i=1}^{S} a_i = A\}} \prod_{i=1}^{S} \frac{\bar{s}(n_i, a_i)\bar{s}(a_1, 1)}{\bar{s}(n_i, 1)} \quad (3)$$

where a_i is the number of ancestors of the species i, and the summation is over $a_i = 1, \ldots, n_i$ with the restriction that the a_i sum to A.$\bar{s}(n_i, a_i)$ is defined as:

$$\bar{s}(n_i, a_i) = \sum_{\{D_{+,i} | a_i\}} \left(\frac{n_i!}{\prod_{j=1}^{n_i} J^{\phi_{i,j}} \phi_{i,j}!} \right) \quad (4)$$

and $\bar{s}(n_i, 1)$ and $\bar{s}(a_i, 1)$ are factorials of $(n_i - 1)$ and $(a_i - 1)$, respectively (Tavaré and Ewens, 1997).

Then we used the following equation to compare the observed community and neutral theory predicted community:

$$D = -2 \ln\left(\frac{L_0}{L_1}\right) = -2[\ln(L_0) - \ln(L_1)] \quad (5)$$

where L_0 represents the log-likelihood of the null model and L_1 represents the log-likelihood of the alternative model, and D is the deviation. The *p-value* was computed *via* an X^2-distribution with the degree of freedom being one.

Etienne (2005) sampling formula is used to test the neutrality of fecal microbial communities through Etienne's *Exact* test of neutrality. The Etienne's "Exact neutrality test," which is based on the sequential construction schemes, does not require alternative model in hypothesis testing. Therefore, it avoids the discussion of validity of the alternative model in empirical evaluations (Etienne, 2007). In brief, firstly, we apply the maximum likelihood estimation (MLE) method to estimate the parameters of the neutral model. This process was performed using the R package UNTB (available at: https://cran.r-project.

org/web/packages/untb/index.html). Secondly, for each sample, we simulated 100 artificial communities (datasets) using the estimated parameters (θ, I, J) and then calculated the likelihood for each artificial dataset via Etienne formula, namely P_s. Finally, we compared the mean of the likelihoods (P_s) of 100 artificial datasets for each sample and the likelihood (P_0) of the corresponding observed sample using a Chi-squared test. The null hypothesis is that there is no significant difference between the probability from the observed community and the values computed from the artificial data sets. If no significant difference between P_s and P_0 were detected, the community would be judged as neutral. The p-value of 0.05 ($p > 0.05$) is adopted as the threshold for passing the neutrality test.

Sloan's (2006) Neutral Theory Model

Sloan et al. (2006) derived an alternative neutral model based on Hubbell's (2001) neutral theory. Sloan's model was aimed to address the difficulty in inferring the taxa-abundance distribution of a microbial community from small metagenomic samples. Sloan's model assumes that the local (or destination) community is saturated with a total of N_T individuals. In the local community, an individual either dies locally or immigrate from the remote (source) community, which occurs at a species-independent rate δ. An immigrant from a source community, with probability m, would immediately replace the dead individual, or a local-born member with probability 1–m would replace it. Hence, the destination community is assembled/reassembled (formed and developed) through a continuous cycle of immigration, reproduction and death. Further assuming that deaths are uniformly distributed in time, then one death is expected during a period of time $1/\delta$. In the meantime, the i-th species, whose initial absolute abundance was N_i, would either increase by one, stay the same or decrease by one with the probability specified by the following three expressions, respectively.

$$\Pr(N_i + 1/N_i) = \left(\frac{N_T - N_i}{N_T}\right)\left[mp_i + (1-m)\left(\frac{N_i}{N_T - 1}\right)\right] \quad (6)$$

$$\Pr(N_i/N_i) = \frac{N_i}{N_T}\left[mp_i + (1-m)\left(\frac{N_i - 1}{N_T - 1}\right)\right] + \left(\frac{N_T - N_i}{N_T}\right)\left[m(1-p_i) + (1-m)\left(\frac{N_T - N_i - 1}{N_T}\right)\right] \quad (7)$$

$$\Pr(N_i - 1/N_i) = \frac{N_i}{N_T}\left[m(1-p_i) + (1-m)\left(\frac{N_T - N_i}{N_T - 1}\right)\right] \quad (8)$$

Let x_i be the *occurrence frequency* of the i-th species in the destination community, i.e., $x_i = n/N$,where n is the number of local community samples where species i occurred and N is the total number of local community samples (Burns et al., 2016), p_i is the occurrence frequency of i-th species in the source community, i.e., the counterpart of x_i in the destination community Sloan et al. (2006) showed that x_i should follow the following beta distribution:

$$x_i \sim Beta[N_T m p_i, N_T m(1-p_i)]. \quad (9)$$

Specifically,

$$\phi_i(x_i; N_T, p_i, m) = c x_i^{N_T m p_i - 1}(1-x_i)^{N_T m(1-p_i)-1}, \quad (10)$$

$$c = \frac{\Gamma(N_T m)}{\Gamma[N_T m(1-p_i)]\Gamma(N_T m p_i)}, \quad (11)$$

where N_i and N_T are the total number of individuals of species i and the total number of individuals (of all species) in the local community samples, respectively, m is the migration frequency, and ϕ_i represents the probability density function, rather than the number of species mentioned in Equation 1.

According to Burns et al. (2016), the process for testing Sloan et al. (2006) neutral model can be summarized as the following three steps.

(*i*) Compute p_i and x_i, with both p_i and x_i, one can fit the beta distribution (Equations. 7, 8) and obtain the estimation of m.

(*ii*) Compute the predicted (theoretical) φ_i the theoretical occurrence frequency of species i across all destination community samples, based on m and the beta distribution (Equation 8).

(*iii*) Judge whether or not the observed x_i of species i falls within its theoretical interval φ_i predicted from the neutral model, and obtain a list of neutral species whose observed x_i satisfy the prediction from the neutral model.

Unlike Hubbell (2001) neutral theory model, there is not a community level statistic (p-value) for testing neutrality with Sloan's model (Sloan et al., 2006, 2007), other than the percentage of neutral or non-neutral species. Obviously, it is not easy to define what "majority" level of the neutral species to designate the whole community as neutral as in the case of Hubbell's model. Another important metric that can be utilized to judge the goodness-of-fitting for Sloan's model is the R^2 or R-squared, the coefficient of determination. Another important metric that can be utilized to judge the goodness-of-fitting for Sloan's model is the R^2 or R-squared, the coefficient of determination. We use a subjective threshold of R-squared = 0.5 as passing the Sloan model test.

RESULTS AND DISCUSSION

Testing the Influence of Obesity on Neutrality at the Community Level

We tested the neutrality of gut microbial community samples using Etienne sampling formula. The model parameters were estimated using the MLE (maximum likelihood estimation) & LLR (log-likelihood ratio) test, as detailed in Etienne (2005, 2007) and Li and Ma (2016). To perform the LLR test, we compared the log-likelihood of each observed gut microbial community with the average log-likelihood of corresponding simulated communities based on the neutral model, and the p-value of the LLR test was listed in the online **Supplementary Table S1**.

The results in **Supplementary Table S1** show that there were only 2 gut microbial communities (subject ID: TS75.2_298948 and TS98_299220) out of 283 communities that passed Etienne neutrality test of Hubbell's neutral theory. Both the communities

satisfying the neutral community model were sampled from the obese patients, and their neutral model parameters are summarized in the following **Table 1**. **Figure 1** displays the graphs of fitting the neutral theory model to these two communities that passed the neutrality exact test.

The test results presented in **Supplementary Table S1** and **Table 1**, as well as **Figure 1**, revealed that, at the whole community level, the number of communities (only 2 out of 283) passing the neutrality test of Hubbell's neutral theory is negligible. Therefore, the assembly processes of gut microbiota should be dominantly shaped by host environmental effects rather than by stochastic neutral effects such as birth/death stochasticities. While the compositions and diversities of gut microbial communities may be different between obese and healthy people as demonstrated in existing studies (Ley et al., 2006; Liu et al., 2017; Menni et al., 2017), obesity is not strong enough to change the intrinsic mechanisms of the community assembly and diversity maintenance in the gut microbiome. In other words, the structure of gut microbiome is primarily shaped by rather strong deterministic host environment, and stochasticities in gut microbial communities do not play a significant role in shaping the assembly of gut microbiome. Furthermore, obesity as a relatively common health disorder nevertheless, does not change the landscape of gut microbiome assembly.

Testing the Influence of Obesity on Neutrality at the Species Level

While the previous section was focused on testing the influence of obesity on gut microbiota neutrality at the whole community level based on Hubbell's (2001) neutral model, here our focus is the neutrality at species level based on Sloan et al. (2006, 2007) neutral model. Because the results from testing Sloan's neutral model may be influenced by samples sizes, we randomly sampled 50 microbiota samples from the lean and obese treatments, respectively, to achieve balanced sample sizes between both the treatments. We further repeated this sampling process 30 times. The averages of the 30 times were taken as the final results of testing Sloan's neutral model (**Table 2**) and the standard deviations were displayed in **Supplementary Table S2**.

The parameters listed in **Table 2** included the average individuals in destination community (N), the immigration rate (m), the goodness-of-fitting (R^2), and the total number of species in each treatment (*Total*). The column "*Neutral*" in **Table 2** listed the percentage of the species within the 95% confidence intervals predicted by the best-fitted neutral model. These species

TABLE 1 | The gut microbial communities passing the neutrality exact test with Etienne sampling formula based on (Hubbell, 2001) neutral theory model*.

Treatment	ID	*J*	*S*	θ	*m*	Log(L_0)	Log(L_1)	*q*-value	*p*-value
Obese	TS75.2_298948	1676	148	38.947	0.99997	−86.334	−85.809	1.051	0.3052
	TS98_299220	2602	177	42.752	0.99991	−110.238	−108.549	3.379	0.0660

**The total number of reads (total individuals) in the sample community (J), the number of species (S), the fundamental biodiversity (θ), the immigration probability (m), log-likelihood of the observed sample [log(L_0)], log-likelihood predicted by the neutral model (log(L_1)), and the log-likelihood ratios (q-value and p-value). P-value >0.05 indicates the community satisfies the prediction of Hubbell's neutral theory.*

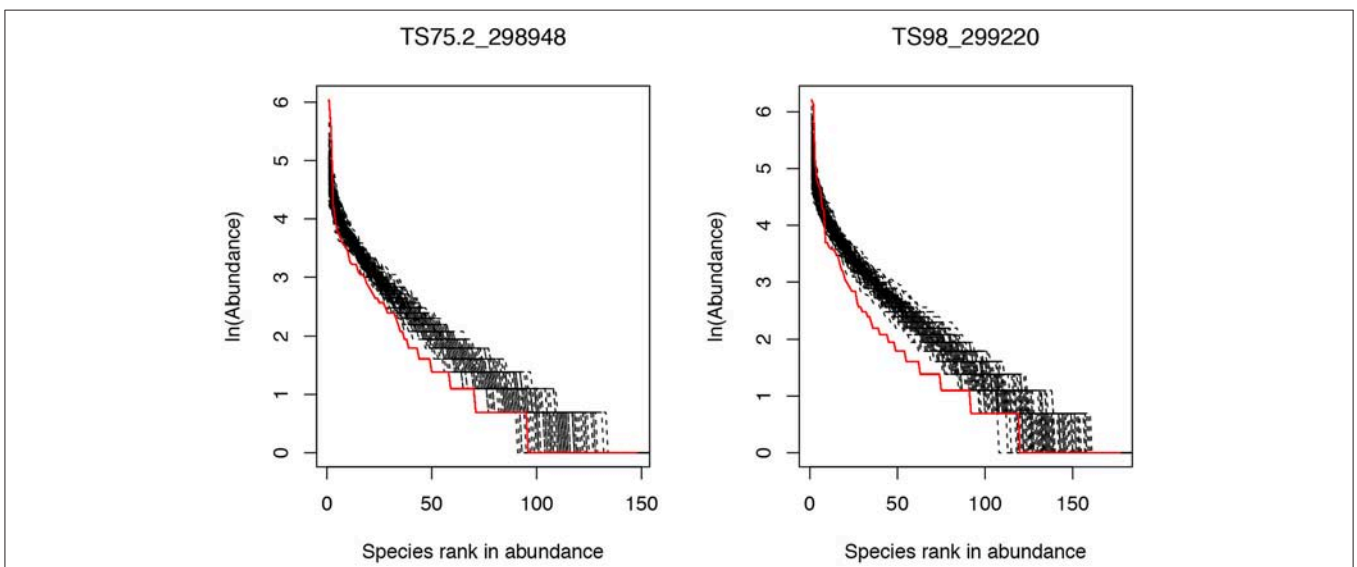


FIGURE 1 | The rank abundance curves of two community samples that successfully passed the neutrality test: the solid red line represents for the observed community and the black dash lines for the simulated communities based on the neutral theory model. The *X*-axis is the species rank order in abundance and *Y*-axis is the abundance of each species in natural logarithm.

follow Sloan's neutral theory. The column "*Non-neutral*" listed the percentage of the species deviating from the prediction of Sloan's neutral model.

As shown in **Table 2**, there are 65.5 and 68.5% of the species that satisfied Sloan's neutral theory in the gut microbial communities of the lean and obese treatment, respectively. In other words, in more than a half of the species in the gut microbiome, stochastic neutral effects are significant. In addition, there were no significant differences in the percentage of neutral species between the obesity and lean treatments (t-test: $p > 0.05$, **Figure 2**). We also tested Sloan's neutral model by treating the lean treatment as source community and the obese treatment as the destination community, and the percentage of neutral species is slightly less (58.6%) than those of the lean or obese treatment alone.

The results from testing Sloan's neutral model seemed to be in conflict with the results from testing Hubbell's neutral model. Why are there more than a half of neutral species in a non-neutral community? The apparent contradiction can be easily resolved if we recall that Hubbell's neutral theory is tested at the whole community level, and a portion of the non-neutral species in a community is sufficient to change the behavior of the whole community. Since Sloan's model tests the neutrality of individual species, theoretically, only if all species in a community pass Sloan's neutrality test, then it should be guaranteed that the whole community is neutral in terms of Hubbell's model. In our study, there were still more than 1/3 of species that clearly demonstrated non-neutral behavior, hence, the results from both the neutral models not only do not contradict with each other, but also present complementary insights for understanding the community assembly mechanisms of the human gut microbiome.

Conclusions and Discussion

In summary, in this study, we applied both Hubbell's and Sloan's neutral theory models to test the influence of obesity on the gut microbiome assembly from both community and species perspectives. At community level, we found that all 283 but 2 gut microbial community samples we tested failed to pass the test of Hubbell's neutral theory, and obesity did not affect the test results. We conclude that the gut microbiome, as a whole, is not neutral and is governed by deterministic host effects. Obesity does not play a significant role in determining the rules (mechanisms) of gut microbiome assembly. From a species perspective, although more than a half of the species in gut microbiome were neutral according to Sloan's neutral model,

TABLE 2 | The gut microbial species passing the test of Sloan's neutral theory in the gut microbiome of lean and obese treatments*.

Source	Destination	*N*	*m*	R^2	Total	Neutral (%)	Non-neutral (%)
Lean	Lean	3629	0.043	0.416	1640	65.5	34.5
Obese	Obese	2543	0.063	0.472	1476	68.5	31.5
Lean	Obese	2586	0.032	0.296	1220	58.6	41.4

**N is the average individuals in destination community, m is the immigration probability, R^2 is the goodness-of-fitting, total is the total number of species in the treatment, neutral is the percentage of the species within the 95% confidence interval predicted by the neutral model, and non-neutral is the percentage of the species deviating from the neutral model.*

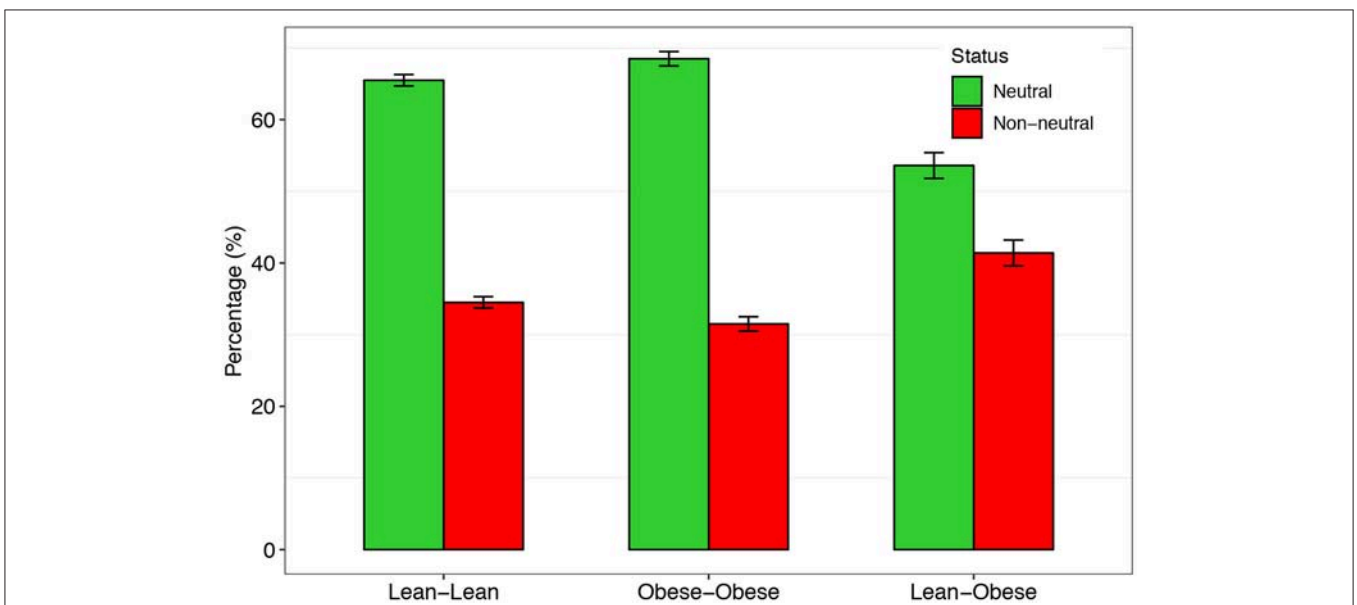


FIGURE 2 | The percentages of the neutral (green) and non-neutral (red) species, respectively, in the three regimes designed for testing Sloan's neutral theory.

it is the minority (∼1/3 of species) that ultimately determined the behavior' of community as a whole. Our findings suggest that gut microbial community is a world consisting of both neutral and non-neutral species, whose collective behavior (i.e., assembly and diversity maintenance mechanisms) is determined by the non-neutral ones. Furthermore, we failed to detect a significant influence of the obesity on neutrality at either species or community scale.

Testing the neutral theory models has been challenging, at least, because of the following four factors: (*i*) the availability of quality data, (*ii*) the availability of computationally efficient algorithms, (*iii*) the neutral model itself, and (*iv*) the interpretation of the test results. First, ideally, the datasets should be sampled from a metacommunity setting consistent with the model assumption, but in practice, such datasets are not easy to obtain. Second, fitting the neutral models with a truly multi-site setting (allowing the computation of variable migration rates among different local communities) was challenging until Harris et al. (2015) recent work, who developed an efficient machine-learning based algorithm. Nevertheless, the adoption of their fitting approach has been slow, possibly due to the availability of suitable datasets. For example, the datasets used in this study and Harris et al. (2015) approach cannot be utilized to test the neutral theory because we cannot assume there are exchanges of microbes (migrations) among individual subjects in ecological time and the neutral theory is largely an ecological time-scale model. Third, obviously, the neutrality assumption is overly simplified, and more recent niche-neutral hybrid models (e.g., Tang and Zhou, 2013) can help to determine the relative significance of deterministic niche forces vs. stochastic neutral forces. Yet, among the four challenges (factors), the most challenging task is to accurately interpret the results from fitting the neutral or niche-neutral hybrid models. For example, it has been suggested that neutral theory can help to determine the significance of drift, dispersal, and speciation, the three of the four key processes for driving community dynamics (the other is selection) (Vellend, 2010; Rosindell et al., 2011, 2012). The difficulty lies in the fact that processes such as dispersal may not be stochastic and instead may be asymmetric among species. In other words, dispersal may be an adaptive behavior in many cases. Therefore, to accurately interpret the results from neutrality test, additional mechanistic studies should be conducted. That said, our study has significant room to improve given the previous discussed challenges. To fully understand the mechanisms of gut microbiome assembly as well as the influences of obesity on the mechanisms, additional biomedical studies including manipulative experiments with animal models should be performed. Nevertheless, we believe that the cross-scale approach we adopted in this study should also be helpful for addressing those challenges.

AUTHOR CONTRIBUTIONS

SM designed the study and wrote the paper. WL, YY, and YX performed the data analysis and interpretations. YS and YM participated in the data interpretation and discussion. All authors approved the submission.

FUNDING

The research received funding from the following sources: NSFC (National Science Foundation of China) (Grant No. 71473243), Yun-Ridge Industrial Talent Grant, a China-US International Collaboration Grant from Yunnan Province, China, and a Grant for Supporting Excellent Undergraduate Internship from Chinese Academy of Sciences.

ACKNOWLEDGMENTS

We appreciate the computational discussion with LW Li and J Li of Kunming Institute of Zoology, Chinese Academy of Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02320/full#supplementary-material

REFERENCES

Alonso, D., Etienne, R. S., and McKane, A. J. (2006). The merits of neutral theory. *Trends Ecol. Evol.* 21, 451–457. doi: 10.1016/j.tree.2006.03.019

Armougom, F., Henry, M., Vialettes, B., Raccah, D., and Raoult, D. (2009). Monitoring bacterial community of human gut microbiota reveals an increase in *Lactobacillus* in obese patients and Methanogens in anorexic patients. *PLoS ONE* 4:e7125. doi: 10.1371/journal.pone.0007125

Burns, A. R., Stephens, W. Z., Stagaman, K., Wong, S., Rawls, J. F., Guillemin, K., et al. (2016). Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME J.* 10, 655–664. doi: 10.1038/ismej.2015.142

Chen, X., Huang, Y., You, Z., et al. (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

Davis, R. A. H., Plaisance, E. P., and Allison, D. B. (2018). Complementary hypotheses on contributors to the obesity epidemic. *Obesity* 26, 17–21. doi: 10.1002/oby.22071

Etienne, R. S. (2005). A new sampling formula for neutral biodiversity: A new sampling formula. *Ecology Letters*, 8, 253–260. doi: 10.1111/j.1461-0248.2004.00717.x

Etienne, R. S. (2007). A neutral sampling formula for multiple samples an 'exact' test for neutrality. *Ecol. Lett.* 10, 608–618. doi: 10.1111/j.1461-0248.2007.01052.x

Fleissner, C. K., Huebel, N., Abd El-Bary, M. M., et al. (2010). Absence of intestinal microbiota does not protect mice from diet-induced obesity. *Br. J. Nutr.* 104, 919–929. doi: 10.1017/S0007114510001303

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

Goodrich, J. K., Davenport, E. R., Waters, J. L., Clark, A. G., and Ley, R. E. (2016b). Cross-species comparisons of host genetic associations with the microbiome. *Science* 352, 532–535. doi: 10.1126/science.aad9379

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

Harris, B. K., Pearsons, T. L., Ijaz, U. Z., Lahti, L., Holmes. I., and Quince C. (2015). Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process. *Proc. IEEE* 105, 516–529. doi: 10.1109/JPROC.2015.2428213

Henao-Mejia, J., Elinav, E., Thaiss, C. A., and Flavell, R. A. (2014). Inflammasomes and metabolic disease. *Annu. Rev. Physiol.* 76, 57–78. doi: 10.1146/annurev-physiol-021113-170324

Hildebrandt, M. A., Hoffmann, C., Sherrill-Mix, S. A., Keilbaugh, S. A., Hamady, M., Chen, Y. Y., et al. (2009). High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology* 137, 1716–1724.e1–2. doi: 10.1053/j.gastro.2009.08.042

Hjorth, M. F., Roager, H. M., Larsen, T. M., et al. (2017). Pre-treatment microbial Prevotella-to-Bacteroides ratio, determines body fat loss success during a 6-month randomized controlled diet intervention. *Int. J. Obes.* 1–4. doi: 10.1038/ijo.2017.220

Huang, Y. A., You, Z. H., Chen, X., Huang, Z. A., Zhang, S., and Yan, G. Y. (2017b). 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

Huang, Z. A., Chen, X., Zhu, Z., Liu, H., Yan, G. Y., You, Z. H., et al. (2017a). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol*. 8:233. doi: 10.3389/fmicb.2017.00233

Hubbell, S. P. (1997). A unified neutral theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs* 16, S9–S21.

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton, NJ: Princeton University Press.

Le Chatelier, E., 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. doi: 10.1038/nature12506

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *PNAS* 102, 11070–11075. doi: 10.1073/pnas.0504978102

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, L., and Ma, Z. S. (2016). Testing the neutral theory of biodiversity with human microbiome datasets. *Sci. Rep.* 6:31448.doi: 10.1038/srep31448

Liu, R., Hong, J., Xu, X., Feng, Q., Zhang, D., Gu, Y., et al. (2017). Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat. Med.* 23, 859–868. doi: 10.1038/nm.4358

Matthews, T. J., Borges, P. A. V., and Whittaker, R. J. (2014). Multimodal species abundance distributions: a deconstruction approach reveals the processes behind the pattern. *Oikos* 123, 533–544. doi: 10.1111/j.1600-0706.2013.00829.x

Menni, C., Jackson, M., Pallister, T., et al. (2017). Gut microbiome diversity and high-fiber intake are related to lower long-term weight gain. *Int. J. Obes.* 41, 1099–1105. doi: 10.1038/ijo.2017.66

Murphy, E. F., Cotter, P. D., Healy, S., Marques, T. M., O'Sullivan, O., Fouhy, F., et al. (2010). Composition and energy harvesting capacity of the gut microbiota: relationship to diet, obesity and time in mouse models. *Gut* 59, 1635–1642. doi: 10.1136/gut.2010.215665

Ravussin, E., and Ryan, D. H. (2018). Three New Perspectives on the Perfect Storm: What's Behind the Obesity Epidemic? *Obesity* 26, 9–10. doi: 10.1002/oby.22085

Rosindell, J., Hubbell, S. P., and Etienne, R. S. (2011). The unified neutral theory of biodiversity and biogeography at age ten. *Trends Ecol. Evol.* 26, 340–348. doi: 10.1016/j.tree.2011.03.024

Rosindell, J., Hubbell, S. P., and He, F. (2012). The case for ecological neutral theory. *Trends Ecol. Evol.* 27, 203–208. doi: 10.1016/j.tree.2012.01.004

Schwiertz, A., Taras, D., Schäfer, K., Beijer, S., Bos, N. A., Donus, C., et al. (2010). Microbiota and SCFA in lean and overweight healthy subjects. *Obesity* 18, 190–195. doi: 10.1038/oby.2009.167

Sloan, W. T., Lunn, M., Woodcock, S., Head, I. M., Nee, S., and Curtis, T. P. (2006). Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* 8, 732–740. doi: 10.1111/j.1462-2920.2005.00956.x

Sloan, W. T., Woodcock, S., Lunn, M., Head, I. M., and Curtis, T. P. (2007). Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microb. Ecol.* 53, 443–455. doi: 10.1007/s00248-006-9141-x

Stephens, R. W., Arhire, L., and Covasa, M. (2018). Gut microbiota: from microorganisms to metabolic organ influencing obesity. *Obesity* 26, 801–809. doi: 10.1002/oby.22179

Sun, S., Ji, Y., Kersten, S., and Qi, L. (2012). Mechanisms of inflammatory responses in obese adipose tissue. *Annu. Rev. Nutr.* 32, 261–286. doi: 10.1146/annurev-nutr-071811-150623

Tang, J., and Zhou, S. (2013). Hybrid niche-neutral models outperform an otherwise equivalent neutral model for fitting coral reef data. *J. Theor. Biol.* 317, 212–218. doi: 10.1016/j.jtbi.2012.10.019

Tavaré, S., and Ewens, W. J. (1997). "Multivariate ewens distribution, in *Discrete Multivariate Distributions,* eds N. L. Johnson, S. Kotz, and N. Balakrishnan (New York, NY: Wiley), 232–246.

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, 480–484. doi: 10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

Vellend, M. (2010). Conceptual synthesis in community ecology. *Q. Rev.Biol.* 85, 183–206. doi: 10.1086/652373

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Wills, C., Condit, R., Foster, R. B., and Hubbell, S. P. (1997). Strong density- and diversity-related effects help to maintain tree species diversity in a neotropical forest. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1252–1257. doi: 10.1073/pnas.94.4.1252

Zhang, C., Zhang, M., Wang, S., Han, R., Cao, Y., Hua, W., et al. (2010). Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. *ISME J.* 4, 232–241. doi: 10.1038/ismej.2009.112

Zhao, L. (2013). The gut microbiota and obesity: from correlation to causality. *Nat Rev. Microbiol.* 11, 639–647. doi: 10.1038/nrmicro3089

Conflict of Interest Statement: 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.

Copyright © 2018 Li, Yuan, Xia, Sun, Miao and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 01 November 2018
doi: 10.3389/fmicb.2018.02560

Human Microbe-Disease Association Prediction With Graph Regularized Non-Negative Matrix Factorization

Bin-Sheng He[1]*, *Li-Hong Peng*[2]* and *Zejun Li*[3,4]*

[1] The First Affiliated Hospital, Changsha Medical University, Changsha, China, [2] School of Information Engineering, Changsha Medical University, Changsha, China, [3] College of Information Science and Engineering, Hunan University, Changsha, China, [4] School of Computer and Information Science, Hunan Institute of Technology, Hengyang, China

A microbe is a microscopic organism which may exists in its single-celled form or in a colony of cells. In recent years, accumulating researchers have been engaged in the field of uncovering microbe-disease associations since microbes are found to be closely related to the prevention, diagnosis, and treatment of many complex human diseases. As an effective supplement to the traditional experiment, more and more computational models based on various algorithms have been proposed for microbe-disease association prediction to improve efficiency and cost savings. In this work, we developed a novel predictive model of Graph Regularized Non-negative Matrix Factorization for Human Microbe-Disease Association prediction (GRNMFHMDA). Initially, microbe similarity and disease similarity were constructed on the basis of the symptom-based disease similarity and Gaussian interaction profile kernel similarity for microbes and diseases. Subsequently, it is worth noting that we utilized a preprocessing step in which unknown microbe-disease pairs were assigned associated likelihood scores to avoid the possible negative impact on the prediction performance. Finally, we implemented a graph regularized non-negative matrix factorization framework to identify potential associations for all diseases simultaneously. To assess the performance of our model, cross validations including global leave-one-out cross validation (LOOCV) and local LOOCV were implemented. The AUCs of 0.8715 (global LOOCV) and 0.7898 (local LOOCV) proved the reliable performance of our computational model. In addition, we carried out two types of case studies on three different human diseases to further analyze the prediction performance of GRNMFHMDA, in which most of the top 10 predicted disease-related microbes were verified by database HMDAD or experimental literatures.

Keywords: microbe, disease, association prediction, graph regularization, matrix factorization

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Edwin Wang,
University of Calgary, Canada
Jiawei Luo,
Hunan University, China

***Correspondence:**
Bin-Sheng He
hbscsmu@163.com
Li-Hong Peng
plhhnu@163.com
Zejun Li
lzjfox@163.com

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

Received: *16 August 2018*
Accepted: *08 October 2018*
Published: *01 November 2018*

Citation:
He B-S, Peng L-H and Li Z (2018) Human Microbe-Disease Association Prediction With Graph Regularized Non-Negative Matrix Factorization. Front. Microbiol. 9:2560. doi: 10.3389/fmicb.2018.02560

INTRODUCTION

Antonie Van Leeuwenhoek, the father of microbiology, was the first to discover, observe, describe, study, and conduct scientific experiments with microbes, using simple single-lensed microscopes of his own design in 1673 (Leeuwenhoek, 1683-1775). From then on, with the development of biological theory and technology, a great mass of microbes has been discovered. It has been suggested that the amount of organisms living below the Earth's surface is comparable with the amount of life on or above the surface (Gold, 1992). As we know, microbes are very closely related

to humans in many fields, such as food production (Smid and Lacroix, 2013), water treatment (Tabatabaei et al., 2010), energy (Tanaka, 1999), and human health (Thiele et al., 2013). Especially, many studies have demonstrated that one of the most important effects of microbes on humans is the associations between microbes and complex human diseases. For example, Boleij et al. (2015) proved that the *Bacteroides fragilis* toxin gene is associated with colorectal neoplasia, especially in late-stage colorectal cancer (CRC). Moreover, Galiana et al. (2014) found that *Actinomyces* can be as an indicator in the evolution of chronic obstructive pulmonary disease (COPD) patients because their study confirmed a strong association between the presence or absence of *Actinomyces* and the severity of the clinical condition. Another example is that periodontal pathogens *Porphyromonas gingivalis* and *Fusobacterium nucleatum* stimulate tumorigenesis of oral squamous cell carcinoma (OSCC) via direct interaction with oral epithelial cells through Toll-like receptors which is beneficial to the development of corresponding prevention and treatment schemes (Binder Gallimidi et al., 2015). Thus, due to the fact that detecting potential microbiological markers could help to provide a better understanding of the pathogenesis of diseases and the role played by the microbiota in its severity, it is of great significance to explore the potential associations between microbes and diseases. However, since traditional experimental methods always suffer from the time constraints and capital limitations, proposing novel computational models is able to be an effective complement for uncovering potential microbe-disease associations. Recently, many feasible and effective prediction models have been developed by researchers.

In the last few years, some prediction models were proposed based on network analysis. Ma et al. (2017) developed an analysis method based on the microbe-based human disease network (Human Microbe Disease Network, HMDN) to infer the associations between microbes and disease genes, symptoms, chemical fragments, and drugs. In the method, they first utilized a large-scale text mining-based method to build the microbe-disease association network, on which the cosine similarity was calculated for each disease pair to construct the HMDN. Taking microbe-disease gene association prediction as an example, the potential related disease genes of a microbe in the HMDN can be finally obtained by finding the highly overlapped genes among the microbe-related diseases in the gene-based human disease network (Human Gene Disease Network, HGDN). Besides, in a similar way, this analysis method can also be used between HMDN and symptom-based human disease network (Human Symptoms Disease Network, HSDN), chemical fragment-based human disease network (Human Chemical Fragments Disease Network, HCDN), and drug-based human disease network (Human Drug Disease Network, HDDN) to infer the associations between microbes and disease symptoms, chemical fragments, and drugs, respectively. However, the prediction performance of this analysis method is limited by the small microbe-based disease network. Thereafter, Chen et al. (2017a) was the first to propose a computational model of KATZ measure for Human Microbe-Disease Association prediction (KATZHMDA) on a large scale. Firstly, they integrated the known microbe-disease associations network and Gaussian interaction profile kernel similarity networks of microbes and diseases into a heterogeneous graph. Through summarizing all walks with different weighted lengths (i.e., the walk with shorter length was assigned larger coefficient) for each microbe-disease pair, they finally calculated the association probability between each microbe and disease. Moreover, KATZHMDA is applicable for new diseases/microbes without known associations if there are additional available similarity information between the new disease/microbe and other diseases/microbes in the known microbe-disease association network. One limitation of KATZHMDA is that the optimal value of the number of walks is still hard to select. Later,Huang Z. A. et al. (2017) proposed a model of Path-Based Human Microbe-Disease Association Prediction (PBHMDA) by integrating known microbe-disease association network and Gaussian interaction profile kernel similarity network for microbes and diseases into a heterogeneous interlinked network in which a threshold was set to remove the edges that represent weak correlations. In the heterogeneous interlinked network, the weights of all paths between a microbe-disease pair were finally aggregated to represent the association probability between the microbe and the disease, while the weight of each path was calculated by multiplying the weights of all edges in the path without overlap and then penalizing the path with a decay coefficient. The limitation existing in PBHMDA is that it will cause bias to microbes or diseases with more known associations. Moreover, PBHMDA cannot work well for new microbes and new diseases.

In addition, some proposed models were not based on network analysis. Since the negative microbe-disease samples (i.e., microbe-disease pairs that are confirmed to have no associations) are unavailable, Wang et al. (2017) presented a semi-supervised learning-based computational model of Laplacian Regularized Least Squares for Human Microbe-Disease Association prediction (LRLSHMDA) by optimizing the Laplacian regularized least squares classifiers in microbe space and disease space. Finally, they used a simple weighted average operation on the above two optimal classifiers to obtain the final probability matrix that indicates the potential association probabilities between microbes and diseases. However, LRLSHMDA is still faced with the problem of being unable to be implemented to new diseases without known associated microbes. Similarly, with no need for negative samples, Huang Y. A. et al. (2017) developed the method of a Neighbor- and Graph-based combined Recommendation model for Human Microbe-Disease Association prediction (NGRHMDA) by combining two recommendation models that are neighbor-based collaborative filtering model and topological information-based model. In the neighbor-based collaborative filtering model, considering that different microbe-disease pairs may share the same microbes or diseases, they computed two association possibility matrices respectively from the microbe perspective and disease perspective and then averaged them to obtain a prediction matrix. While in the topological information-based model, they introduced a two-step diffusion approach on the microbe-disease bipartite graph to obtain another prediction matrix. Ultimately, the above two prediction matrices

were simply averaged to get the final association possibilities for all microbe-disease pairs. What is worth noting is that NGRHMDA shares the same aforementioned disadvantage with LRLSHMDA.

In summary, all of the above models have their own limitations in predicting microbe-disease associations. Due to the lack of measurements for microbe/disease similarity, some models are only based on the Gaussian interaction profile kernel similarity of microbes and diseases that leads to unavoidable bias to those well-investigated diseases and microbes. Besides, some models cannot predict for new microbes/diseases and optimal parameters in some models are not easy to select. In this work, considering some of the above limitations, we developed a novel computational model of Graph Regularized Non-negative Matrix Factorization for Human Microbe-Disease Association prediction (GRNMFHMDA). First of all, the information of Gaussian interaction profile kernel similarity of microbes and diseases, symptom-based disease similarity and known microbe-disease associations in HMDAD (Ma et al., 2017) were combined as the input to start the whole prediction process. Here, after data preparation, the prediction process consists of two main steps, the preprocessing step and the step of GRNMF. In the preprocessing step, the weighted K nearest neighbor profiles for microbes and diseases were calculated to reconstruct the original adjacency matrix obtained based on the known microbe-disease associations so that we could avoid the possible negative impact on the final prediction performance from unknown microbe-disease pairs. While in the step of GRNMF, Tikhonov (L_2) and graph Laplacian regularization were introduced into the standard NMF framework to obtain a smoother solution from matrix factorization and take full advantage of the geometric structure of our data, respectively. In addition, global leave-one-out cross validation (LOOCV), local LOOCV and two types of case studies were carried out to evaluate the prediction performance of our model. As a result, GRNMFHMDA obtained AUCs of 0.8715 (global LOOCV) and 0.7898 (local LOOCV). More than that, 9 (Asthma), 9 (Obesity), and 8 (Type 1 diabetes) out of the top 10 predicted disease-related microbes were confirmed by HMDAD or experimental literatures. Thus, it is obvious that our model would perform well in microbe-disease association prediction according to the aforementioned results.

MATERIALS AND METHODS

Method Overview

Here, to predict potential associations between microbes and diseases, the model of GRNMFHMDA (See **Figure 1**) can be decomposed into three steps: (1) data preparation, in which adjacency matrix, microbe similarity, and disease similarity were established; (2) the preprocessing step, in which unknown microbe-disease pairs were assigned with associated likelihood scores based on the calculation of weighted K nearest neighbor profiles for microbes and diseases; (3) GRNMF, in which Tikhonov (L_2) and Graph Laplacian regularization were introduced into the standard NMF framework to obtain the final score matrix.

Human Microbe-Disease Associations

From the Human Microbe-Disease Association Database (HMDAD, http://www.cuilab.cn/hmdad) (Ma et al., 2017), we can download 483 known microbe-disease associations between 292 microbes and 39 human diseases. However, since some microbe-disease associations we downloaded are the same, there were only 450 known associations after removing the duplicate parts according to different evidences. In order to represent the associations information in a more convenient and efficient way, we defined an adjacency matrix $Y \in R^{m*n}$, where m and n denoted the number of microbes and diseases, respectively. Moreover, the element $Y(m_i, d_j)$ was set to 1 if microbe m_i and disease d_j had known association, otherwise 0.

Gaussian Interaction Profile Kernel Similarity for Microbes

There is a hypothesis that similar microbes (i.e., microbes exhibiting a similar pattern of interaction and non-interaction with the diseases of a microbe-disease association network) are inclined to be associated with the same disease, on which many previous studies had relied to construct the Gaussian interaction profile kernel similarity for microbes (Chen et al., 2017a; Huang Z. A. et al., 2017). In this article, based on the same assumption, we first represent the interaction profile for each microbe with a binary vector involving the association information between the microbe and each disease in the known microbe-disease association network. On the basis of the definition of adjacency matrix Y, the ith row vector ($Y(m_i) = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$) can be used to denote the interaction profile of microbe m_i. Thus, according to the method of van Laarhoven et al. (2011), the Gaussian interaction profile kernel similarity between microbe m_i and m_j can be defined as follows:

$$S^m(m_i, m_j) = \exp(-\gamma_m \left\| Y(m_i) - Y(m_j) \right\|^2) \quad (1)$$

where

$$\gamma_m = \gamma'_m / \left(\frac{1}{m} \sum_{i=1}^{m} \left\| Y(m_i) \right\|^2 \right) \quad (2)$$

Here, γ_m is the adjustment coefficient that can be obtained by normalizing another bandwidth parameter γ'_m.

Gaussian Interaction Profile Kernel Similarity for Diseases

The construction of the Gaussian interaction profile kernel similarity for diseases is based on the assumption that similar diseases (i.e., diseases exhibiting a similar pattern of interaction and non-interaction with the microbes of a microbe-disease association network) are more likely to be associated with similar microbes. Here, the interaction profile for each disease is also represented by a binary vector containing the association information between the disease and each microbe in the known microbe-disease association network. Based on the same method of van Laarhoven et al. (2011), the jth column vector ($Y(d_j) =$

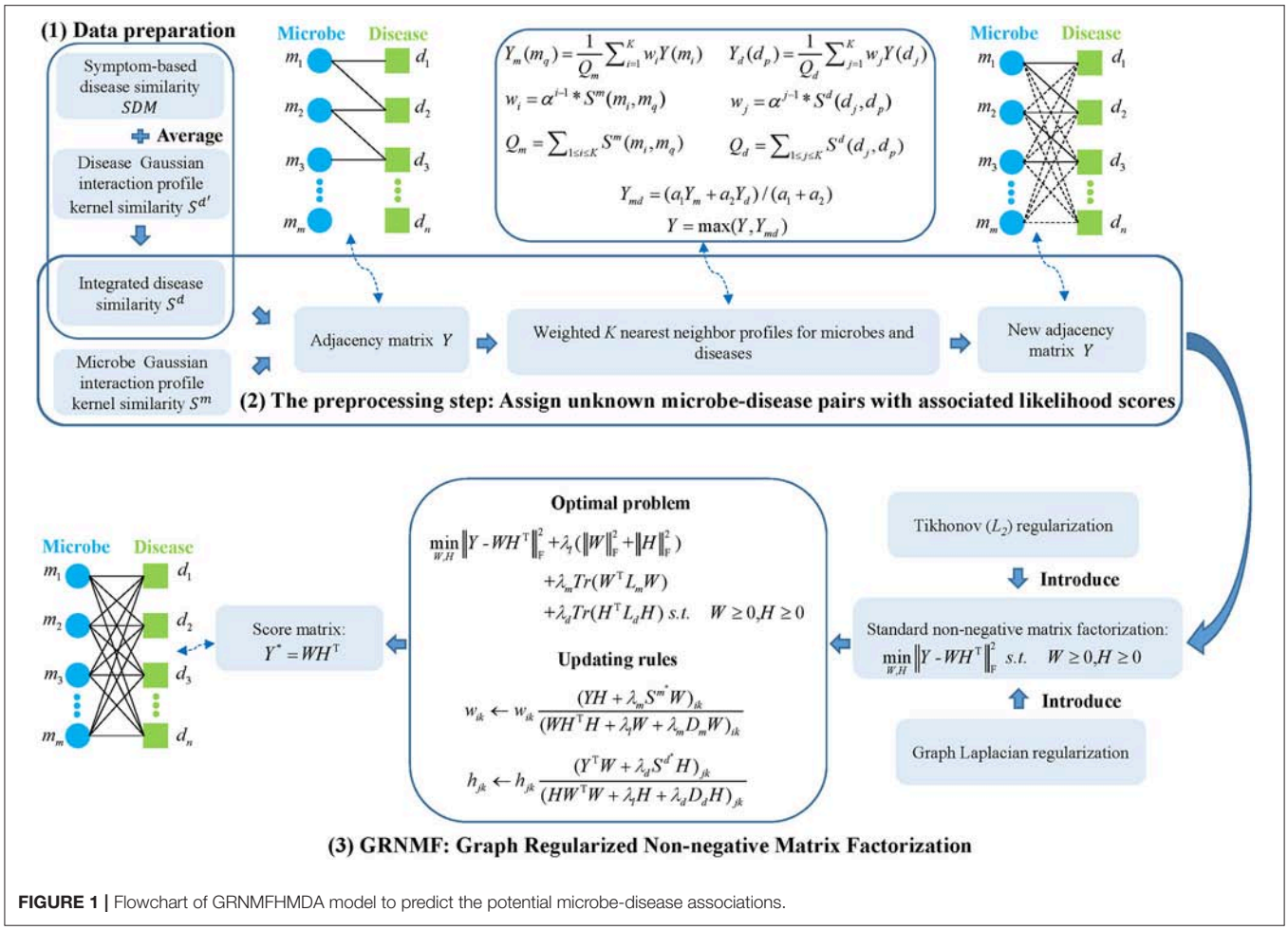


FIGURE 1 | Flowchart of GRNMFHMDA model to predict the potential microbe-disease associations.

$(Y_{1j}, Y_{2j}, \ldots, Y_{mj})$) denotes the interaction profile of disease d_j and the Gaussian interaction profile kernel similarity between disease d_i and d_j can be defined as follows:

$$S^{d'}(d_i, d_j) = \exp(-\gamma_d \| Y(d_i) - Y(d_j) \|^2) \quad (3)$$

where

$$\gamma_d = \gamma'_d / \left(\frac{1}{n} \sum_{j=1}^{n} \| Y(d_j) \|^2 \right) \quad (4)$$

Similarly, γ_d is the adjustment coefficient that can be calculated by normalizing another bandwidth parameter γ'_d.

Integrated Symptom-Based Disease Similarity

As we have mentioned above, Gaussian interaction profile kernel similarity is used in our model to measure the similarity of microbes and diseases. However, since the Gaussian interaction profile kernel similarity is an association information-based measurement, it is essential to combine more types of microbe or disease similarities based on other available biological information. Indeed, according to different biological data, many researchers have developed their own method to measure the similarity of microbes or diseases. For instance, Zhou et al. (2014) proposed a model of symptom-based human disease network (HSDN) to measure the disease similarity based on co-occurrence of disease/symptom terms recorded in different literatures. In this work, we implemented HSDN to calculate symptom-based disease similarity (*SDM*) and then constructed a new disease similarity matrix (S^d) by integrating *SDM* with $S^{d'}$ in an average way according to the study of Chen et al. (2017a):

$$S^d = \frac{S^{d'} + SDM}{2} \quad (5)$$

Weighted *K* Nearest Neighbor Profiles for Microbes and Diseases

Due to the fact that values in interaction profiles of microbes or diseases without known associations are all zeros, the prediction performance may be affected to some extent. Considering that, to deal with the above mentioned problem, we came up with a preprocessing step to establish new interaction profiles both for

microbes and diseases. For each microbe m_q, we first find out its K nearest known microbes, each of which must has at least one known association. Next, the similarity information between m_q and its K nearest known microbes together with the information of their corresponding K interaction profiles are combined to calculate the new interaction profile as follows:

$$Y_m(m_q) = \frac{1}{Q_m} \sum_{i=1}^{K} w_i Y(m_i) \tag{6}$$

where

$$w_i = \alpha^{i-1*} S^m(m_i, m_q) \tag{7}$$

$$Q_m = \sum_{1 \le i \le K} S^m(m_i, m_q) \tag{8}$$

Here, m_1 to m_K denote the K nearest known microbes of m_q which were sorted in descending order based on the similarity values between them. The function of the weight coefficient w_i is that the corresponding similarity value is assigned higher weight if m_i is more similar to m_q. Besides, α is a decay term whose value is in the range of [0,1] and Q_m is the normalization term.

In a similar way, the new interaction profile for each disease d_p can be defined as follows:

$$Y_d(d_p) = \frac{1}{Q_d} \sum_{j=1}^{K} w_j Y(d_j) \tag{9}$$

$$w_j = \alpha^{j-1*} S^d(d_j, d_p) \tag{10}$$

$$Q_d = \sum_{1 \le j \le K} S^d(d_j, d_p) \tag{11}$$

After calculating the new interaction profiles from microbe perspective and disease perspective, we combine Y_m and Y_d as follows:

$$Y_{md} = (a_1 Y_m + a_2 Y_d)/(a_1 + a_2) \tag{12}$$

where a_1 and a_2 are two weight coefficient and both of them are set to 1 for simplicity.

Finally, to replace the element $Y(m_i, d_j) = 0$ with an associated likelihood score, we use the following equation to update the original adjacency matrix Y.

$$Y = \max(Y, Y_{md}) \tag{13}$$

GRNMF

As a common method, the purpose of the standard NMF is to find two non-negative matrices whose product is an optimal approximation to the original matrix (Sotiras et al., 2015; Xu et al., 2015). Therefore, the adjacency matrix $Y \in R^{m*n}$ can be decomposed into two parts after implementing NMF, namely, $W \in R^{m*k}$ and $H \in R^{n*k}$ ($Y \approx WH^{\mathrm{T}}$). Accordingly, we can further get the following standard optimization problem:

$$\min_{W,H} \left\| Y - WH^{\mathrm{T}} \right\|_{\mathrm{F}}^2 + L(W, H) \tag{14}$$

where $L(W, H)$ is a regularization term to prevent overfitting.

Here, motivated by the study of Xiao et al. (2017) and the standard NMF framework, we introduced other two terms, the Tikhonov (L_2) (Guan et al., 2011) and graph Laplacian regularization (Cai et al., 2011), to predict microbe-disease associations. The utilizing of Tikhonov regularization aims to obtain a smooth solution (W and H), while the purpose of introducing graph regularization is to ensure a part-based representation through taking full advantage of the data geometric structure. Thus, we can construct the optimization problem of GRNMF as follows:

$$\begin{aligned} &\min_{W,H} \left\| Y - WH^{\mathrm{T}} \right\|_{\mathrm{F}}^2 + \lambda_l(\|W\|_{\mathrm{F}}^2 + \|H\|_{\mathrm{F}}^2) \\ &+\lambda_m \sum_{i,p=1}^{n} \left\| w_i - w_p \right\|^2 S^{m*}{}_{ip} + \lambda_d \sum_{j,q=1}^{m} \left\| h_j - h_q \right\|^2 S^{d*}{}_{jq} \; s.t. \; W \ge 0, H \ge 0 \end{aligned} \tag{15}$$

Here, λ_l, λ_m and λ_d are the corresponding regularization coefficients. Besides,w_i and h_j are defined as ith rows of W and j th rows of H, respectively. In order to avoid negative affects to the prediction performance of our model, we introduced sparse weight matrices of S^{d*} and S^{m*}that are constructed on the basis of the geometrical information of disease and microbe data spaces (S^d and S^m), respectively. Then, Equation (14) can be transformed into:

$$\begin{aligned} &\min_{W,H} \left\| Y - WH^{\mathrm{T}} \right\|_{\mathrm{F}}^2 + \lambda_l(\|W\|_{\mathrm{F}}^2 + \|H\|_{\mathrm{F}}^2) \\ &+\lambda_m Tr(W^T L_m W) + \lambda_d Tr(H^{\mathrm{T}} L_d H) \; s.t. \; W \ge 0, H \ge 0 \end{aligned} \tag{16}$$

where $Tr(\bullet)$ represents the trace of a matrix. Here, L_m and L_d are the corresponding graph Laplacian matrices for S^{m*} and S^{d*} that can be calculated as follows:

$$L_m = D_m - S^{m*} \tag{17}$$

$$L_d = D_d - S^{d*} \tag{18}$$

where D_m and D_d are the diagonal matrices whose entries are row (or column) sums of S^{m*} and S^{d*}, respectively.

Based on the information of the nearest neighbor graph on a scatter of data points, researchers came up with a conclusion that local geometric structure is able to be effectively modeled (Cai et al., 2011; Li et al., 2017). Since microbes or diseases appearing in the same cluster are more likely to behave similarly, according to the above conclusion, we construct the graph matrices S^{m*} and S^{d*} in terms of microbe space and disease space respectively on the basis of the p nearest neighbors and corresponding clustering information. Here, we use the ClusterONE method (Nepusz et al., 2012) to construct the graph S^{m*} from microbe space, in which the

weight matrix X^m is generated based on the microbe similarity matrix S^m as follows:

$$X_{ij}^m = \begin{cases} 1\ i \in N(m_j)\ \&\ j \in N(m_i), m_j \in C \\ 0\ i \notin N(m_j)\ \&\ j \notin N(m_i), m_j \notin C \\ 0.5\ otherwise \end{cases} \quad (19)$$

where $N(m_i)$ and $N(m_j)$ are the sets of p nearest neighbors of m_i and m_j, respectively. C denotes to any one of the clusters obtained by ClusterONE method and we define the graph matrix S^{m^*} for microbes as follows:

$$\forall i, j\ S_{ij}^{m^*} = X_{ij}^m S_{ij}^m \quad (20)$$

In a similar way as the computation of S^{m*}, we calculate the graph matrix S^{d^*} according to the disease similarity matrix S^d.

Here, we defined Φ=[φ_{ik}] and Ψ=[ψ_{jk}] as the Lagrange multipliers for the constrains $w_{ik} \geq 0$ and $h_{jk} \geq 0$, respectively. In this work, we first convert the optimization problem in Equation (15) to an unconstraint problem, then minimize this problem by utilizing the corresponding Lagrange function L_f as follows:

$$\begin{aligned} L_f &= Tr(YY^{\mathrm{T}}) - 2Tr(YHW^{\mathrm{T}}) + Tr(WH^{\mathrm{T}}HW^{\mathrm{T}}) + \lambda_l Tr(WW^{\mathrm{T}}) \\ &+ \lambda_l Tr(HH^{\mathrm{T}}) + \lambda_m Tr(W^{\mathrm{T}}L_m W) + \lambda_d Tr(H^{\mathrm{T}}L_d H) \\ &+ Tr(\Phi W^{\mathrm{T}}) + Tr(\Psi H^{\mathrm{T}}) \end{aligned} \quad (21)$$

To solve the above problem, we first calculate the partial derivatives with respect to W and H as follows:

$$\frac{\partial L_f}{\partial W} = -2YH + 2WH^{\mathrm{T}}H + 2\lambda_l W + 2\lambda_m L_m W + \Phi \quad (22)$$

$$\frac{\partial L_f}{\partial H} = -2Y^{\mathrm{T}}W + 2HW^{\mathrm{T}}W + 2\lambda_l H + 2\lambda_d L_d H + \Psi \quad (23)$$

After using the Karush-Kuhn-Tucker (KKT) conditions of $\varphi_{ik}w_{ik} = 0$ and $\psi_{jk}h_{jk} = 0$ (Facchinei et al., 2014), we can obtain the equations for w_{ik} and h_{jk} as follows:

$$\begin{aligned} &- (YH)_{ik}w_{ik} + (WH^{\mathrm{T}}H)_{ik}w_{ik} + (\lambda_l W)_{ik}w_{ik} \\ &+ [\lambda_m(D_m - S^{m^*})W]_{ik}w_{ik} = 0 \end{aligned} \quad (24)$$

$$\begin{aligned} &- (Y^{\mathrm{T}}W)_{jk}h_{jk} + (HW^{\mathrm{T}}W)_{jk}h_{jk} + (\lambda_l H)_{jk}h_{jk} \\ &+ [\lambda_d(D_d - S^{d^*})H]_{jk}h_{jk} = 0 \end{aligned} \quad (25)$$

Finally, on the basis of the above two equations, we can get the updating rules for w_{ik} and h_{jk} as follows:

$$w_{ik} \leftarrow w_{ik}\frac{(YH + \lambda_m S^{m^*}W)_{ik}}{(WH^{\mathrm{T}}H + \lambda_l W + \lambda_m D_m W)_{ik}} \quad (26)$$

$$h_{jk} \leftarrow h_{jk}\frac{(Y^{\mathrm{T}}W + \lambda_d S^{d^*}H)_{jk}}{(HW^{\mathrm{T}}W + \lambda_l H + \lambda_d D_d H)_{jk}} \quad (27)$$

Based on the above two updating formulas, we can obtain the final two non-negative matrices W and H until convergence. Subsequently, we calculate the score matrix Y^* for microbe-disease pairs by utilizing $Y^* = WH^{\mathrm{T}}$, in which the higher score of a microbe-disease pair indicates that the microbe is more likely to be associated with the corresponding disease. In addition, for better understanding, we provided the pseudocode of the whole GRNMF algorithm (See **Figure 2**).

RESULTS

Performance Evaluation

Cross validation, a widely used assessment method, was introduced to evaluate the prediction performance of GRNMFHMDA. In this study, we utilized two types of cross validations, namely, global LOOCV and local LOOCV. For the global LOOCV, each of the known microbe-disease associations was in turn considered to be the test sample while the remaining known associations were treated as the training samples. Besides, all of the unknown microbe-disease pairs were regarded as the candidate samples which would be used in the ranking process. After implementing GRNMFHMDA, we ranked each test sample with all candidate samples according to their predicted scores. As for local LOOCV, the difference is that the test sample was only ranked with the candidate samples involving the investigated disease.

In each cross validation process, we would consider that the test sample was successfully predicted if the ranking of the test sample was higher than the given threshold. Further, based on the ranks of all test samples, we drew a receiver operating characteristic (ROC) curve through calculating the ratio between true positive rate (TPR, sensitivity) and false positive rate (FPR, 1-specificity) under different thresholds both for global LOOCV and local LOOCV. Sensitivity meant the ratio between the number of test samples ranking higher than the given threshold and the number of positive samples (known microbe-disease associations), while 1-specificity denoted the percentage of the number of negative microbe-disease pairs whose ranks were lower than the given threshold. Moreover, area under the ROC curve (AUC) was calculated to make quantitative evaluation for our model's prediction performance. The model would be considered to be able to perfectly predict all associations if the value of AUC equaled to 1, while the model was only supposed to be able to make random prediction if the value of AUC equaled to 0.5. As a result, GRNMFHMDA obtained AUCs of 0.8715 and 0.7898 in global LOOCV and local LOOCV, respectively. Furthermore, the prediction performance of our model outperformed the KATZHMDA both in global LOOCV (0.8644) and local LOOCV (0.6998), which proved the superior accuracy and reliability of our model in predicting microbe-disease associations (See **Figure 3**).

Case Study

Here, we put forward two types of case studies on three different common human diseases with the purpose of further assessing the prediction performance of GRNMFHMDA. On the basis of the known microbe-disease associations in HMDAD,

Algorithm: GRNMF

Input: Matrix $Y \in R^{m*n}$, $S^m \in R^{m*m}$, and $S^d \in R^{n*n}$, decay term α, subspace dimensionality k, neighborhood sizes K and p, regularization coefficients λ_l, λ_m and λ_d

Output: Predicted association matrix Y^*

Initialize: randomly initialize the two non-negative matrices $W \in R^{m*k}$ and $H \in R^{n*k}$

1. calculate weighted K nearest neighbor profiles for microbes:

for each microbe m_q $(q \in [1,m])$ **do**

for $i \leftarrow 1$ to K **do**

$$w_i = \alpha^{i-1} * S^m(m_i, m_q)$$

end for

$$Q_m = \sum_{1 \le i \le K} S^m(m_i, m_q)$$

$$Y_m(m_q) = \frac{1}{Q_m} \sum_{i=1}^{K} w_i Y(m_i)$$

end for

2. calculate weighted K nearest neighbor profiles for diseases:

for each disease d_p $(p \in [1,n])$ **do**

for $j \leftarrow 1$ to K **do**

$$w_j = \alpha^{j-1} * S^d(d_j, d_p)$$

end for

$$Q_d = \sum_{1 \le j \le K} S^d(d_j, d_p)$$

$$Y_d(d_p) = \frac{1}{Q_d} \sum_{j=1}^{K} w_j Y(d_j)$$

end for

3. combine the two weighted K nearest neighbor profiles of microbes and diseases:
$Y_{md} = (a_1 Y_m + a_2 Y_d) / (a_1 + a_2)$

4. update the original adjacency matrix Y:
$Y = \max(Y, Y_{md})$

5. construct matrices S^{m^*} and $\mathbf{S}^{d^*}$ from S^m and $\mathbf{S}^d$, respectively.

6. **while** not converged **do**

$$w_{ik} \leftarrow w_{ik} \frac{(YH + \lambda_m S^{m^*} W)_{ik}}{(WH^{\mathrm{T}} H + \lambda_l W + \lambda_m D_m W)_{ik}}$$

$$h_{jk} \leftarrow h_{jk} \frac{(Y^{\mathrm{T}} W + \lambda_d S^{d^*} H)_{jk}}{(HW^{\mathrm{T}} W + \lambda_l H + \lambda_d D_d H)_{jk}}$$

end while

7. **return** $Y^* = WH^{\mathrm{T}}$

FIGURE 2 | The pseudocode of the whole GRNMF algorithm.

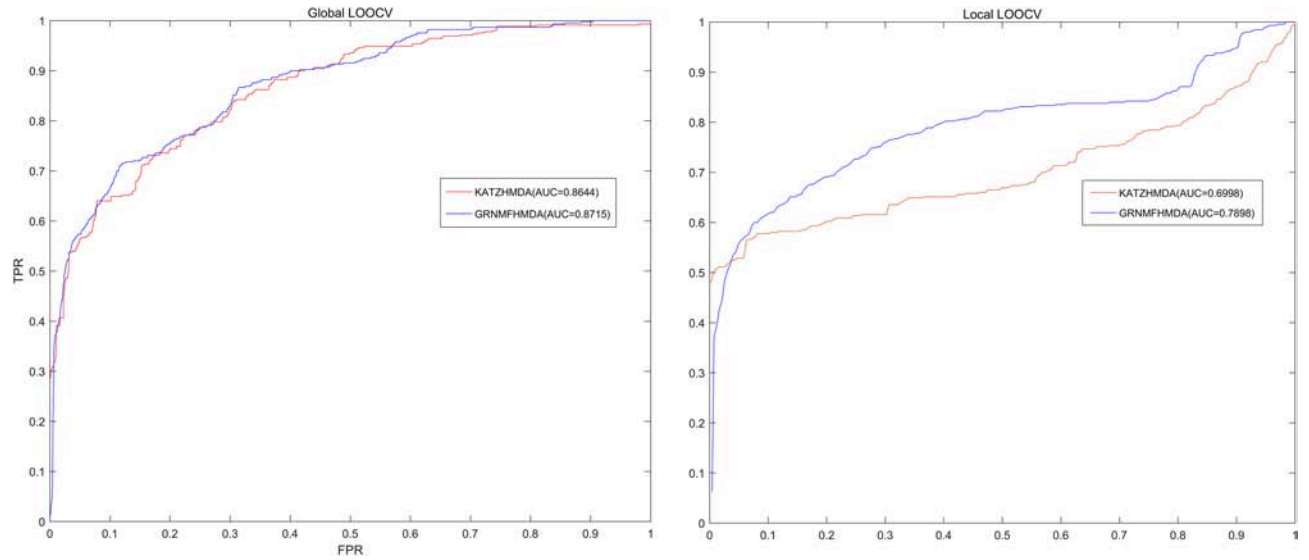


FIGURE 3 | The comparison of prediction performance between GRNMFHMDA and the classical model of KATZHMDA both in global LOOCV and local LOOCV. As a result, GRNMFHMDA achieved AUCs of 0.8715 and 0.7898 in the global and local LOOCV, which exceed the first computational model of KATZHMDA in the field of microbe-disease association prediction.

we implemented GRNMFHMDA to predict disease-related microbes and then validated the top 10 predicted microbes by HMDAD or recent literatures.

Asthma, a common long-term inflammatory disease of the airways of the lungs, often starts during childhood and its average number of deaths and death rates (per 100,000 people) respectively reached to 38 and 0.1 in 2016 in the World Health Organization (WHO) European region among 10–14 years old children (Kyu et al., 2018). Here, under the GRNMFHMDA framework, asthma was treated as an investigated disease to explore its potential associated microbes. As a result, 9 out of the top 10 microbes in the prediction list were confirmed to be associated with asthma by experimental literatures (See **Table 1**). For example, *Lactobacillus casei rhamnosus* Lcr35, a species of *Lactobacillus* (1st in the prediction list), was found to be able to attenuate airway inflammation and hyperreactivity in a mouse model of asthma through oral treatment before sensitization (Yu et al., 2010). Besides, Ding et al. (2018) discovered that exosomes derived by *Pseudomonas* (2nd in the prediction list) *aeruginosa* could induce protection against allergic sensitization in asthma mice. Another example is that there is a distinct alteration of the sputum microbiota with a greater prominence of *Firmicutes* (4th in the prediction list) in severe asthma (Zhang et al., 2016).

Obesity, a medical condition in which accumulated excess body fat reaches a certain level that may have a negative effect on health, is a leading preventable cause of death worldwide (Reinier and Chugh, 2015). In recent years, plenty of studies have shown certain associations between obesity and microbes that helps a lot to the prevention and treatment of obesity. For instance, many researchers have demonstrated that *methanogens* play a specific role in weight gain and the development of obesity in human subjects (Armougom et al., 2009; Krajmalnik-Brown et al., 2012). Not only that, many studies have now been conducted into the potential of probiotics to ameliorate obesity and diabetes (Delzenne et al., 2011; Peterson et al., 2015). Therefore, taking obesity as another investigated disease in the first type of case study, we found that 9 out of the top 10 predicted obesity-related microbes were confirmed by experimental literatures (See **Table 2**). For the phylum *Proteobacteria* (1st in the prediction list) which belongs to gram-negative bacteria, the existing study already discovered that it was abundant in the obese group compared with lean group (Park et al., 2015). Besides, as a species of *Clostridia* (2nd in the prediction list), the presence of *Clostridium ramosum* in simplified human intestinal (SIHUMI) enhanced diet-induced obesity according to the experiment data of Woting et al. (2014). Moreover, *Bacillus*, a genus of *Clostridia* (3rd in the prediction list), was found to have outgrown dramatically in the obesity group by Gao et al. (2018).

TABLE 1 | Prediction list of the top 10 potential asthma-related microbes based on the known associations in HMDAD database and the corresponding validation evidences (experimental literatures in PubMed) for these associations.

Rank	Microbe	Evidence
1	Lactobacillus	PMID: 20592920
2	Pseudomonas	PMID: 29795208
3	Burkholderia	unconfirmed
4	Firmicutes	PMID: 27078029
5	Actinobacteria	PMID: 23265859
6	Clostridium coccoides	PMID: 21477358
7	Streptococcus	PMID: 17928596
8	Clostridia	PMID: 22047069
9	Lachnospiraceae	PMID: 26512904
10	Fusobacterium	PMID: 24024497

TABLE 2 | Prediction list of the top 10 potential obesity-related microbes based on the known associations in HMDAD database and the corresponding validation evidences (experimental literatures in PubMed) for these associations.

Rank	Microbe	Evidence
1	Proteobacteria	PMID: 25407880
2	Clostridia	PMID: 25271283
3	Bacilli	PMID: 29280312
4	Faecalibacterium prausnitzii	PMID: 19849869
5	Clostridium	PMID: 23645850
6	Betaproteobacteria	PMID: 29312822
7	Clostridium coccoides	PMID: 23147032
8	Lactobacillus	PMID: 23056479
9	Fusobacterium nucleatum	unconfirmed
10	Prevotella	PMID: 21695273

TABLE 3 | Prediction list of the top 10 potential type 1 diabetes-related microbes via removing all the known type 1 diabetes-microbe associations in HMDAD database.

Rank	Microbe	Evidence
1	Clostridia	confirmed by HMDAD
2	Proteobacteria	confirmed by HMDAD
3	Clostridium coccoides	unconfirmed
4	Lactobacillus	confirmed by HMDAD
5	Bacteroidetes	confirmed by HMDAD
6	Firmicutes	confirmed by HMDAD
7	Faecalibacterium prausnitzii	PMID: 23934614
8	Clostridium	PMID: 23433344
9	Betaproteobacteria	unconfirmed
10	Bacilli	PMID: 24930037

The validation evidences denote to whether the predicted associations were confirmed by the HMDAD database or experimental literatures in PubMed.

More than that, in order to facilitate future researchers to study the disease-related microbes that they are interested in, based on the known associations in HMDAD, we provided the whole prediction list including all pairs between 292 microbes and 39 diseases as well as their predicted association scores (See **Supplementary Table 1**).

In addition, to prove the predictive applicability of our model on new diseases without known associated microbes, we carried out another case study on a disease via removing all its known associations in HMDAD. In this way, the prediction process of seeking the investigated disease-related microbes can only depend on the information of other known microbe-disease associations (training samples) and the relevant similarity measures. What needs to be emphasized is that only candidate samples (all microbe-disease pairs including the investigated disease) were ranked and then verified in HMDAD. Hence, there was no overlap between training samples and prediction list. In other words, the verification of predicted associations was independent of HMDAD. Type 1 diabetes, a form of diabetes mellitus, is believed to involve a combination of genetic and environmental factors such as dietary agents (Serena et al., 2015), viral infections (Rewers and Ludvigsson, 2016) and gut microbiota (Bibbò et al., 2017). Especially in gut microbiota, the previous study confirmed that the genus *Bacteroides* is the largest representative of type 1 diabetes-associated dysbiosis that can be modulated by diet (Mejjía-León and Barca, 2015). Thus, considering the significance of studying type 1 diabetes-related microbes, we took type 1 diabetes as the investigated disease to predict its potential associated microbes under the framework of the second type of case study. After implementing GRNMFHMDA, we obtained the ranks of type 1 diabetes' candidate microbes in terms of their association scores (See **Table 3**). As a result, 8 out of the top 10 predictions were confirmed by HMDAD or recent literatures. For example, Giongo et al. (2011) demonstrated that the *Clostridia* (1st in the prediction list) sequences increased in control samples (samples of general population) as the abundance of *Clostridia* decreased overtime in the case samples (samples of patients with type 1 diabetes). Moreover, at the phylum level and at p-values $<$ 0.001, *Proteobacteria* (2nd in the prediction list) was found to be higher in case samples than that in control samples (Brown et al., 2011). Another example is that *Lactobacillus strains* (a species of *Lactobacillus* ranking 4th in the prediction list) was found to be able to induce specific changes in the immune system of non-obese diabetic (NOD) mice that can increase or decrease diabetes (Brown et al., 2011).

According to the results presented, GRNMFHMDA consistently achieved an excellent predictive performance in the two types of case studies. With the continuous experimental research on microbe-disease associations, we expect that more and more microbes in the prediction lists generated by our model would be verified in the future.

DISCUSSION

In this article, we proposed a novel prediction model of GRNMFHMDA based on the known microbe-disease associations in HMDAD, Gaussian interaction profile kernel similarity of microbes and diseases and symptom-based disease similarity. To eliminate the possible problem caused by unknown microbe-disease pairs that may affect our final prediction performance, we first implemented a preprocessing step to establish new interaction profiles both for microbes and diseases. Then, after introducing Tikhonov (L_2) and graph Laplacian regularization under the standard NMF framework, we finally obtained reliable and satisfactory prediction performance both in LOOCV and case studies. Therefore, we can conclude that our prediction model is able to play critical role in revealing the associations between microbes and diseases, thus improving the prevention, diagnosis and treatment of many complex human diseases in the future.

Here, the reason why GRNMFHMDA performed well in microbe-disease association prediction lies in the following facts. Firstly, in the study of Wang et al. (2015), to model cancer hallmark traits and networks, nodes and links in the network were weighted, and certain scoring functions were developed to

represent gene regulatory logics/strengths on networks. Inspired by that, based on the data extracted from the acknowledged databases, we implemented proper and effective measurements to quantify microbe-disease association network, microbe similarity network and disease similarity network, which guaranteed the reliable prediction performance of our model. Secondly, before implementing GRNMF, we constructed new interaction profiles both for microbes and diseases to further assign those unknown microbe-disease pairs with associated likelihood score, which also improved our model's performance in some degree. Thirdly, different from the standard NMF, we introduced Tikhonov (L_2) and graph Laplacian regularization that ensured the final two non-negative matrices smoothness and guaranteed a part-based representation via fully exploiting the data geometric structure, respectively.

Nevertheless, here are also some limitations restricting the accuracy of our model that need to be overcome in future studies. Initially, the types of similarities for microbes and diseases are not enough yet and we believe that our model would be significantly improved with more biological data and similarity measurements being taken into consideration. Successful advance in association prediction research in various fields of computational biology would also accelerate the development of effective models for microbe-disease association prediction (Chen and Yan, 2013; Chen et al., 2016, 2017b, 2018a,b,c; Chen and Huang, 2017; You et al., 2017). Secondly, as shown in the research of Hao et al. (2018), three representative genome-scale cellular networks, genome-scale metabolic network (GMN), transcriptional regulatory network (TRN), and signal transduction network (STN), were found to be able to become a necessary tool in the systematic analysis of microbes through network integration. Therefore, whether there are similar molecular networks between two microbes is well worth studying in constructing our prediction model. Thirdly, the selection of the optimal parameters is still worth studying. Finally, GRNMFHMDA would inevitably cause bias to diseases that have more known associated microbes and vice versa. Hence, we would come up with optimization strategies to deal with those limitations in our next work.

AUTHOR CONTRIBUTIONS

B-SH conceived the project, developed the prediction method, analyzed the result, and revised the paper. L-HP designed the experiments, implemented the experiments, analyzed the result, and wrote the paper. ZL analyzed the result and revised the paper. All authors read and approved the final manuscript.

FUNDING

B-SH was supported by Key Program of Hunan Provincial Education Department (Grant No. 15A026), General Program of Hunan Provincial Philosophy and Social Science Planning Fund office (Grant No. 15YBA035). L-HP was supported by Natural Science Foundation of Hunan province under Grant No. 2018JJ3570. ZL was supported by National Nature Science Foundation of China (Grant No. 61672223).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02560/full#supplementary-material

REFERENCES

Armougom, F., Henry, M., Vialettes, B., Raccah, D., and Raoult, D. (2009). Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and Methanogens in anorexic patients. *PLoS ONE* 4:e7125. doi: 10.1371/journal.pone.0007125

Bibbò, S., Dore, M. P., Pes, G. M., Delitala, G., and Delitala, A. P. (2017). Is there a role for gut microbiota in type 1 diabetes pathogenesis? *Ann. Med.* 49, 11–22. doi: 10.1080/07853890.2016.1222449

Binder Gallimidi, A., Fischman, S., Revach, B., Bulvik, R., Maliutina, A., Rubinstein, A. M., et al. (2015). Periodontal pathogens *Porphyromonas gingivalis* and *Fusobacterium nucleatum* promote tumor progression in an oral-specific chemical carcinogenesis model. *Oncotarget* 6, 22613–22623. doi: 10.18632/oncotarget.4209

Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., et al. (2015). The *Bacteroides fragilis* toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin. Infect. Dis.* 60, 208–215. doi: 10.1093/cid/ciu787

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

Cai, D., He, X., Han, J., and Huang, T. S. (2011). graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1548–1560. doi: 10.1109/TPAMI.2010.231

Chen, X., and Huang, L. (2017). LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2017a). 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

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016). NLLSS: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning. *PLoS Comput. Biol.* 12:e1004975. doi: 10.1371/journal.pcbi.1004975

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018a). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* doi: 10.1093/bioinformatics/bty503. [Epub ahead of print].

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018b). BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2017b). MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinformatics*. doi: 10.1093/bib/bbx130. [Epub ahead of print].

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Delzenne, N. M., Neyrinck, A. M., Bäckhed, F., and Cani, P. D. (2011). Targeting gut microbiota in obesity: effects of prebiotics and probiotics. *Nat. Rev. Endocrinol.* 7, 639–646. doi: 10.1038/nrendo.2011.126

Ding, F. X., Liu, B., Zou, W. J., Li, Q. B., Tian, D. Y., and Fu, Z. (2018). *Pseudomonas aeruginosa*-derived exosomes ameliorates allergic reactions via inducing the Treg response in asthma. *Pediatr. Res.* 84, 125–133. doi: 10.1038/s41390-018-0020-1

Facchinei, F., Kanzow, C., and Sagratella, S. (2014). Solving quasi-variational inequalities via their KKT conditions. *Math. Program.* 144, 369–412. doi: 10.1007/s10107-013-0637-0

Galiana, A., Aguirre, E., Rodriguez, J. C., Mira, A., Sañtibanez, M., Candela, I., et al. (2014). Sputum microbiota in moderate versus severe patients with COPD. *Eur. Respir. J.* 43, 1787–1790. doi: 10.1183/09031936.00191513

Gao, R., Zhu, C., Li, H., Yin, M., Pan, C., Huang, L., et al. (2018). Dysbiosis signatures of gut microbiota along the sequence from healthy, young patients to those with overweight and obesity. *Obesity* 26, 351–361. doi: 10.1002/oby.22088

Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., Novelo, L. L., Casella, G., et al. (2011). Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J.* 5, 82–91. doi: 10.1038/ismej.2010.92

Gold, T. (1992). The deep, hot biosphere. *Proc. Natl. Acad. Sci. U.S.A.* 89, 6045–6049. doi: 10.1073/pnas.89.13.6045

Guan, N., Tao, D., Luo, Z., and Yuan, B. (2011). Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE Trans. Image Process.* 20, 2030–2048. doi: 10.1109/TIP.2011.2105496

Hao, T., Wu, D., Zhao, L., Wang, Q., Wang, E., and Sun, J. (2018). The genome-scale integrated networks in microorganisms. *Front. Microbiol.* 9:296. doi: 10.3389/fmicb.2018.00296

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

Huang, Z. A., Chen, X., Zhu, Z., Liu, H., Yan, G. Y., You, Z. H., et al. (2017). PBHMDA: Path-Based Human Microbe-Disease Association Prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Krajmalnik-Brown, R., Ilhan, Z. E., Kang, D. W., and Dibaise, J. K. (2012). Effects of gut microbes on nutrient absorption and energy regulation. *Nutr. Clin. Pract.* 27, 201–214. doi: 10.1177/0884533611436116

Kyu, H. H., Stein, C. E., Boschi Pinto, C., Rakovac, I., Weber, M. W., Dannemann Purnat, T., et al. (2018). Causes of death among children aged 5-14 years in the WHO European Region: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Child Adolesc. Health* 2, 321–337. doi: 10.1016/S2352-4642(18)30095-6

Leeuwenhoek, A. V. (1683-1775). Part of a Letter from Mr Antony van Leeuwenhoek, F. R. S. concerning Green Weeds Growing in Water, and Some Animalcula Found about Them. *Philos. Trans.* 23, 1304–1311.

Li, X., Cui, G., and Dong, Y. (2017). Graph regularized non-negative low-rank matrix factorization for image clustering. *IEEE Trans. Cybern.* 47, 3840–3853. doi: 10.1109/TCYB.2016.2585355

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe-disease associations. *Brief. Bioinformatics* 18, 85–97. doi: 10.1093/bib/bbw005

Mejjía-León, M. E., and Barca, A. M. (2015). Diet, microbiota and immune system in type 1 diabetes development and evolution. *Nutrients* 7, 9171–9184. doi: 10.3390/nu7115461

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* 9, 471–472. doi: 10.1038/nmeth.1938

Park, H. J., Lee, S. E., Kim, H. B., Isaacson, R. E., Seo, K. W., and Song, K. H. (2015). Association of obesity with serum leptin, adiponectin, and serotonin and gut microflora in beagle dogs. *J. Vet. Intern. Med.* 29, 43–50. doi: 10.1111/jvim.12455

Peterson, C. T., Sharma, V., Elmén, L., and Peterson, S. N. (2015). Immune homeostasis, dysbiosis and therapeutic modulation of the gut microbiota. *Clin. Exp. Immunol.* 179, 363–377. doi: 10.1111/cei.12474

Reinier, K., and Chugh, S. S. (2015). Obesity and sudden death: visceral response? *Heart* 101, 165–166. doi: 10.1136/heartjnl-2014-306921

Rewers, M., and Ludvigsson, J. (2016). Environmental risk factors for type 1 diabetes. *Lancet* 387, 2340–2348. doi: 10.1016/S0140-6736(16)30507-4

Serena, G., Camhi, S., Sturgeon, C., Yan, S., and Fasano, A. (2015). The role of gluten in celiac disease and type 1 diabetes. *Nutrients* 7, 7143–7162. doi: 10.3390/nu7095329

Smid, E. J., and Lacroix, C. (2013). Microbe-microbe interactions in mixed culture food fermentations. *Curr. Opin. Biotechnol.* 24, 148–154. doi: 10.1016/j.copbio.2012.11.007

Sotiras, A., Resnick, S. M., and Davatzikos, C. (2015). Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization. *Neuroimage* 108, 1–16. doi: 10.1016/j.neuroimage.2014.11.045

Tabatabaei, M., Rahim, R. A., Abdullah, N., Wright, A. D. G., Shirai, Y., Sakai, K., et al. (2010). Importance of the methanogenic archaea populations in anaerobic wastewater treatments. *Process Biochem.* 45, 1214–1225. doi: 10.1016/j.procbio.2010.05.017

Tanaka, K. (1999). From the fryer to the fuel tank: the complete guide to using vegetable oil as an alternative fuel. *Green Teacher* 15, 46.

Thiele, I., Heinken, A., and Fleming, R. M. (2013). A systems biology approach to studying the role of microbes in human health. *Curr. Opin. Biotechnol.* 24, 4–12. doi: 10.1016/j.copbio.2012.10.001

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Wang, E., Zaman, N., Mcgee, S., Milanese, J. S., Masoudi-Nejad, A., and O'Connor-Mccourt, M. (2015). Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin. Cancer Biol.* 30, 4–12. doi: 10.1016/j.semcancer.2014.04.002

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Woting, A., Pfeiffer, N., Loh, G., Klaus, S., and Blaut, M. (2014). Clostridium ramosum promotes high-fat diet-induced obesity in gnotobiotic mouse models. *MBio* 5, e01530–e01514. doi: 10.1128/mBio.01530-14

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2017). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Xu, J., Xiang, L., Wang, G., Ganesan, S., Feldman, M., Shih, N. N., et al. (2015). Sparse Non-negative Matrix Factorization (SNMF) based color unmixing for breast histopathological image analysis. *Comput. Med. Imaging Graph* 46(Pt 1), 20–29. doi: 10.1016/j.compmedimag.2015.04.002

You, Z. H., Huang, Z. A., Zhu, Z., Yan, G. Y., Li, Z. W., Wen, Z., et al. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005455. doi: 10.1371/journal.pcbi.1005455

Yu, J., Jang, S. O., Kim, B. J., Song, Y. H., Kwon, J. W., Kang, M. J., et al. (2010). The effects of *Lactobacillus rhamnosus* on the prevention of asthma in a murine model. *Allergy Asthma Immunol. Res.* 2, 199–205. doi: 10.4168/aair.2010.2.3.199

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

Zhou, X., Menche, J., Barabási, A. L., and Sharma, A. (2014). Human symptoms-disease network. *Nat. Commun.* 5, 4212. doi: 10.1038/ncomms5212

Conflict of Interest Statement: 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.

Copyright © 2018 He, Peng and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 20 November 2018
doi: 10.3389/fmicb.2018.02762

Evaluating the Value of Defensins for Diagnosing Secondary Bacterial Infections in Influenza-Infected Patients

Siyu Zhou[1], *Xianwen Ren*[2]*, *Jian Yang*[1]* and *Qi Jin*[1]*

[1] MOH Key Laboratory of Systems Biology of Pathogens, Peking Union Medical College, Institute of Pathogen Biology, Chinese Academy of Medical Sciences, Beijing, China, [2] BIOPIC, School of Life Sciences, Peking University, Beijing, China

OPEN ACCESS

Edited by:
Xing Chen,
China University of Mining and Technology, China

Reviewed by:
Zheng Xia,
Oregon Health and Science University, United States
Junjie Yue,
Institute of Biotechnology (CAAS), China

***Correspondence:**
Xianwen Ren
renxwise@pku.edu.cn
Jian Yang
yangj@ipbcams.ac.cn
Qi Jin
zdsys@vip.sina.com

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

Received: 30 August 2018
Accepted: 29 October 2018
Published: 20 November 2018

Citation:
Zhou S, Ren X, Yang J and Jin Q (2018) Evaluating the Value of Defensins for Diagnosing Secondary Bacterial Infections in Influenza-Infected Patients. Front. Microbiol. 9:2762. doi: 10.3389/fmicb.2018.02762

Acute respiratory infections by influenza viruses are commonly causes of severe pneumonia, which can further deteriorate if secondary bacterial infections occur. Although the viral and bacterial agents are quite diverse, defensins, a set of antimicrobial peptides expressed by the host, may provide promising biomarkers that would greatly improve the diagnosis and treatment. We examined the correlations between the gene expression levels of defensins and the viral and bacterial loads in the blood on a longitudinal, precision-medical study of a severe pneumonia patient infected by influenza A H7N9 virus. We found that DEFA5 is positively correlated to the blood load of influenza A H7N9 virus ($r = 0.735$, $p < 0.05$, Spearman correlation). DEFB116 and DEFB127 are positively and DEFB108B and DEFB114 are negatively correlated to the bacterial load. Then the diagnostic potential of defensins to discriminate bacterial and viral infections was evaluated on an independent dataset with 61 bacterial pneumonia patients and 39 viral pneumonia patients infected by influenza A viruses and reached 93% accuracy. Expression levels of defensins in the blood may be of important diagnostic values in clinic to indicate viral and bacterial infections.

Keywords: viral infection, bacterial infection, diagnosis, defensin, gene expression

INTRODUCTION

Acute respiratory infections by influenza viruses are commonly the causes of severe pneumonia, which can further deteriorate if secondary bacterial infections occur (McCullers, 2014). Accurate detection of influenza virus infections and the potential secondary bacterial infections is important to improve the diagnosis and treatment of patients with severe pneumonia. Because the viral and bacterial agents are quite diverse, seeking a broad-spectrum test based on only the characteristics of pathogens is currently still a challenging task. Although the rapidly developed next-generation sequencing (NGS) technology provides a powerful tool to catalog the taxonomic composition of clinical samples, the great technological complexity and high price makes it hard to adopt in clinic soon. Identifying biomarkers that can be readily adopted into clinic is urgently needed. Because different pathogens can result in convergent host responses, identifying broad-spectrum diagnostic biomarkers from the host response is probable. With the rapid development of high-throughput biomedical technologies, the gene expression profiles of host blood can now be readily obtained.

Recently several groups have reported in succession that the gene expression profiles of a certain set of genes in the human blood can robustly discriminate bacterial infections from viral infections and a series of bioinformatics tools have been developed to identify the associations between microbes and host health (Ramilo et al., 2007; Edelman et al., 2009; Zaas et al., 2009; Parnell et al., 2012; Hu et al., 2013; Mejias et al., 2013; Peng et al., 2013; Ye et al., 2014; Suarez et al., 2015; Sweeney et al., 2016; Tsalik et al., 2016; Huang Y. A. et al., 2017; Huang Z. A. et al., 2017; Wang et al., 2017; Chen et al., 2018), suggesting the great potential of host response as the diagnostic signature.

Defensins are diverse members of a large family of antimicrobial peptides that are considered as an important part of the innate immune response of hosts and are found in many

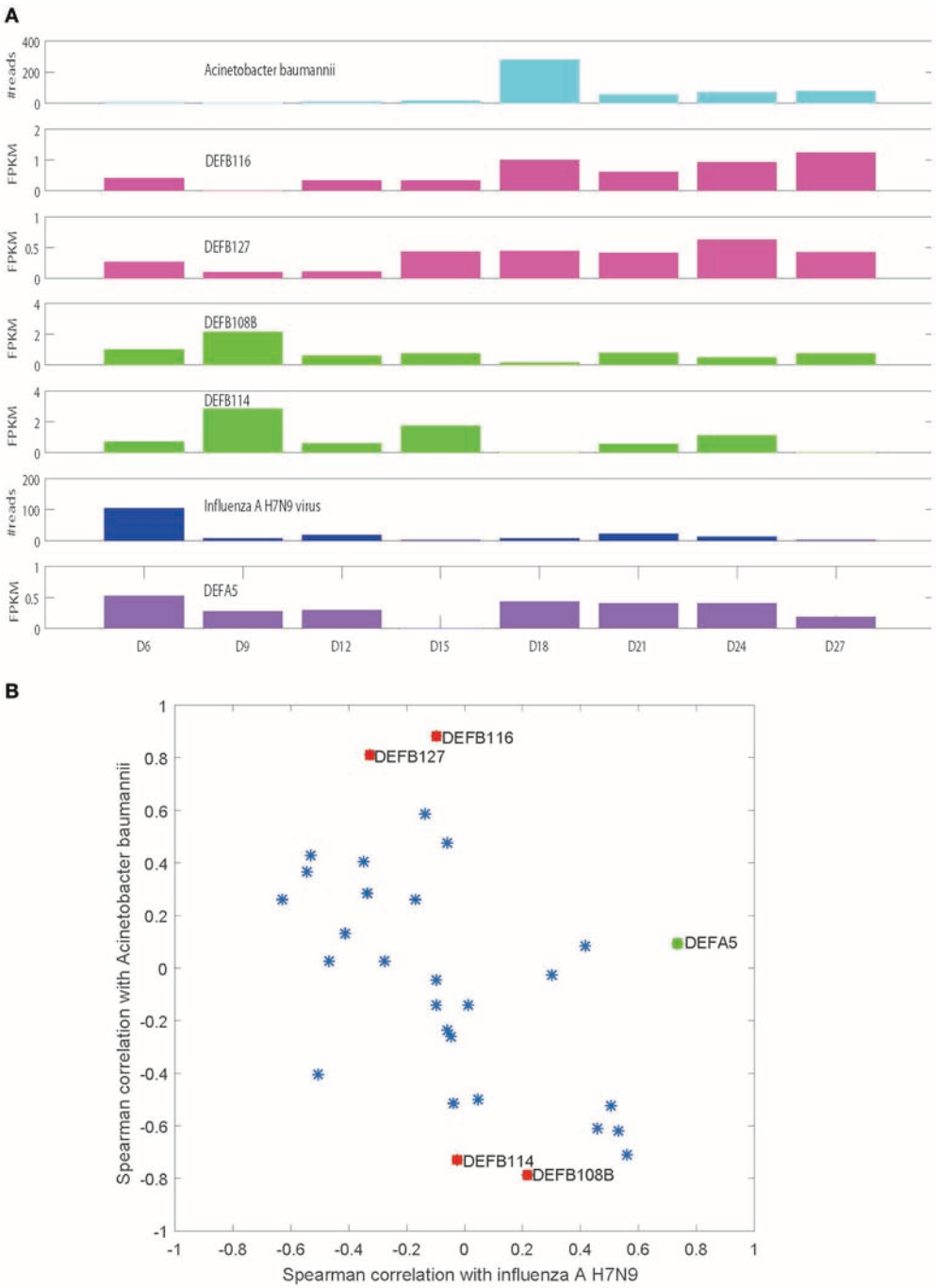


FIGURE 1 | Spearman correlations of defensins and the viral/bacterial loads in blood. **(A)** Plots of the expression levels of selected defensins and the viral/bacterial loads along disease progression. **(B)** Spearman correlations of the total 30 defensins and the viral/bacterial loads.

compartments of the body (Ganz, 2003). These great properties of defensins indicate that they may be good candidates of diagnostic biomarkers to discriminate bacterial/viral infections. However, the currently reported gene signatures identified with human blood gene expression profiles seldom include defensins. It is of pressing need to find out the clinically diagnostic values of defensins.

To reach the objective, we profiled the gene expression levels in blood and the viral and bacterial loads in plasma of a severe pneumonia patient infected by influenza A H7N9 virus via the next-generation sequencing (NGS) technology along with the disease progression. Then we examined the correlations between the expression levels of defensins and the viral and bacterial loads in the blood. Although many defensins did not demonstrate statistically significant correlations with either the viral or the bacterial loads, the *p*-values of several defensins did reach the statistical significance cutoff after multiple-testing corrections. And these statistically significant defensins demonstrated mutually exclusive correlations with the viral loads and the bacterial loads, suggesting that defensins are of great diagnostic values to discriminate viral and bacterial infections. Upon this observation, we then examined the diagnostic potential of defensins on an independent dataset with 61 bacterial pneumonia patients and 39 viral pneumonia patients infected by influenza A viruses (Parnell et al., 2012) via a machine learning method, which confirmed again that defensins are of great diagnostic values to discriminate bacterial infections from viral infections. These results suggest that expression levels of defensins in the blood may be of important diagnostic values in clinic to indicate viral and bacterial infections.

TABLE 1 | The expression levels of the total 30 defensins and the viral/bacterial loads along disease progression.

	Day6	Day9	Day12	Day15	Day18	Day21	Day24	Day27
H7N9	105	6	17	1	6	22	13	1
Acinetobacter baumannii	3	0	7	12	279	54	69	76
DEFB4A	0.00	0.57	0.82	0.58	0.00	0.37	0.19	0.00
DEFB4B	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00
DEFA4	0.00	0.08	0.26	0.57	0.08	0.15	0.08	0.52
DEFA3	0.14	0.11	0.24	0.00	0.00	0.00	0.11	0.00
DEFB131	1.32	3.14	1.13	1.08	0.00	2.04	2.06	0.00
DEFB1	0.46	0.48	0.39	0.25	0.00	0.47	0.47	0.48
DEFA5	0.52	0.28	0.30	0.00	0.44	0.40	0.41	0.18
DEFA6	0.47	0.37	0.67	0.38	0.00	0.12	0.49	0.49
DEFB136	0.90	2.14	0.00	0.73	2.26	0.70	0.00	0.94
DEFB135	2.82	2.24	0.00	0.77	2.36	0.73	0.73	1.97
DEFB133	2.24	0.00	0.00	1.83	0.00	0.00	0.00	2.35
DEFB116	0.40	0.00	0.34	0.32	0.99	0.61	0.92	1.24
DEFB115	0.00	0.48	0.00	0.99	0.00	0.94	1.42	1.91
DEFB114	0.72	2.86	0.62	1.77	0.00	0.56	1.12	0.00
DEFB113	0.00	0.00	1.30	0.62	0.00	1.18	1.19	0.00
DEFB112	0.00	0.24	0.52	1.74	1.27	0.00	0.24	0.00
DEFB110	0.70	0.18	2.20	0.76	0.20	0.36	1.82	2.45
DEFB134	4.95	0.61	4.39	2.10	0.48	0.33	0.34	1.58
DEFB121	0.68	1.24	1.16	0.45	0.60	0.88	0.71	0.34
DEFB132	0.56	0.54	0.51	0.42	0.45	0.35	0.50	0.67
DEFB128	0.75	0.00	0.00	0.41	1.04	0.58	0.00	0.78
DEFB125	0.32	0.26	0.46	0.44	0.27	0.59	0.59	1.02
DEFB123	0.63	0.13	0.41	1.16	0.66	0.37	0.37	0.83
DEFB119	2.59	1.47	0.58	0.82	0.88	0.52	1.34	2.32
DEFB124	1.93	1.80	3.41	1.56	0.66	0.96	4.44	0.00
DEFB108B	1.00	2.15	0.61	0.76	0.17	0.79	0.48	0.75
DEFB118	0.17	0.21	0.52	0.25	0.32	0.27	0.13	0.63
DEFB126	0.09	0.44	0.47	0.52	0.46	0.57	0.21	0.48
DEFB129	0.42	0.66	0.21	0.55	0.14	0.32	0.46	0.53
DEFB127	0.27	0.11	0.11	0.44	0.45	0.41	0.63	0.42

Defensin expression levels were quantified by FPKM and the viral/bacterial loads were quantified by the number of reads.

MATERIALS AND METHODS

Longitudinal Gene Expression Profiles of a Severe Pneumonia Patient Infected by Influenza a H7N9 Virus

The severe pneumonia patient infected by influenza A H7N9 virus was admitted to hospital on Day 5 after illness onset and died on Day 29. Since Day 6, blood samples were collected for every three days, i.e., on Days 6, 9, 12, 15, 18, 21, 24, and 27 after illness onset. The total RNA was isolated and then subjected to sequencing on Illumina Solexa GA II with read length of 80 bp (see Hu et al., 2015 for the technical details). Cufflinks (version 2.1.1, with default parameters) (Trapnell et al., 2010) was used to quantify the gene expression profiles of defensins after mapping the quality-controlled reads to human genome (GRCh37 and Gencode19) using Tophat (version 2.0.10, with default parameters) (Kim et al., 2013). This study was reviewed and approved by the Ethics Committee of the Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College. Written informed consent was obtained for the use of peripheral blood samples from the patient's relatives. This study was carried out in accordance with the recommendations of the Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College. The protocol was approved by the Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

Quantifying the Microbial Species Infecting in the Blood Samples

To quantify the microbial species infecting in the blood samples, a metagenomic analysis method was applied. In detail, the same sequencing reads were aligned to the NCBI non-redundant nucleotide database by BLASTN (version 2.2.22, with parameters "-e 1e-10 –b 10 –v 10") (Altschul et al., 1997). Then, the results were parsed and visualized by the MEGAN software (Huson et al., 2007, 2016; Mitra et al., 2011), upon which those reads specifically mapped to bacterial or viral genomes were counted and exported as the bacterial/viral

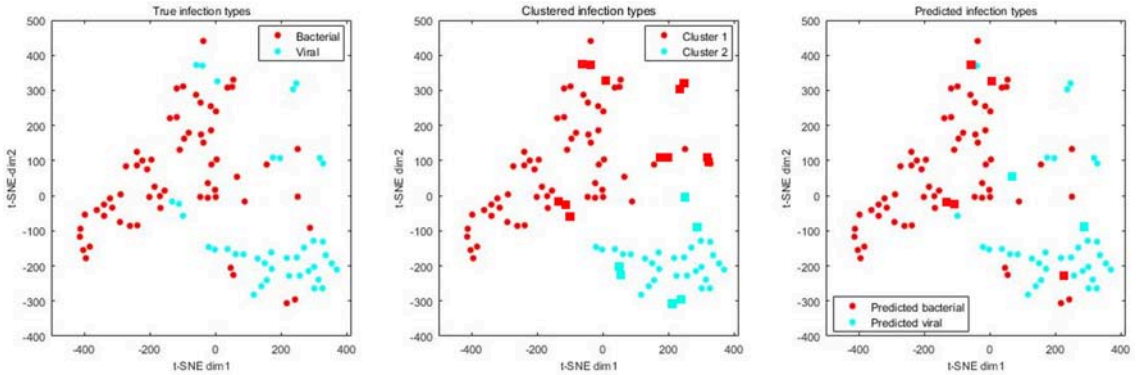


FIGURE 2 | True, clustered and predicted infection types of 61 bacterial and 39 viral pneumonia patients**.** Expression levels of defensins and associated genes were extracted from the whole dataset and then subjected to t-SNE analysis for visualization. Circles mean correctly clustered/classified samples while rectangles mean incorrectly clustered/classified samples.

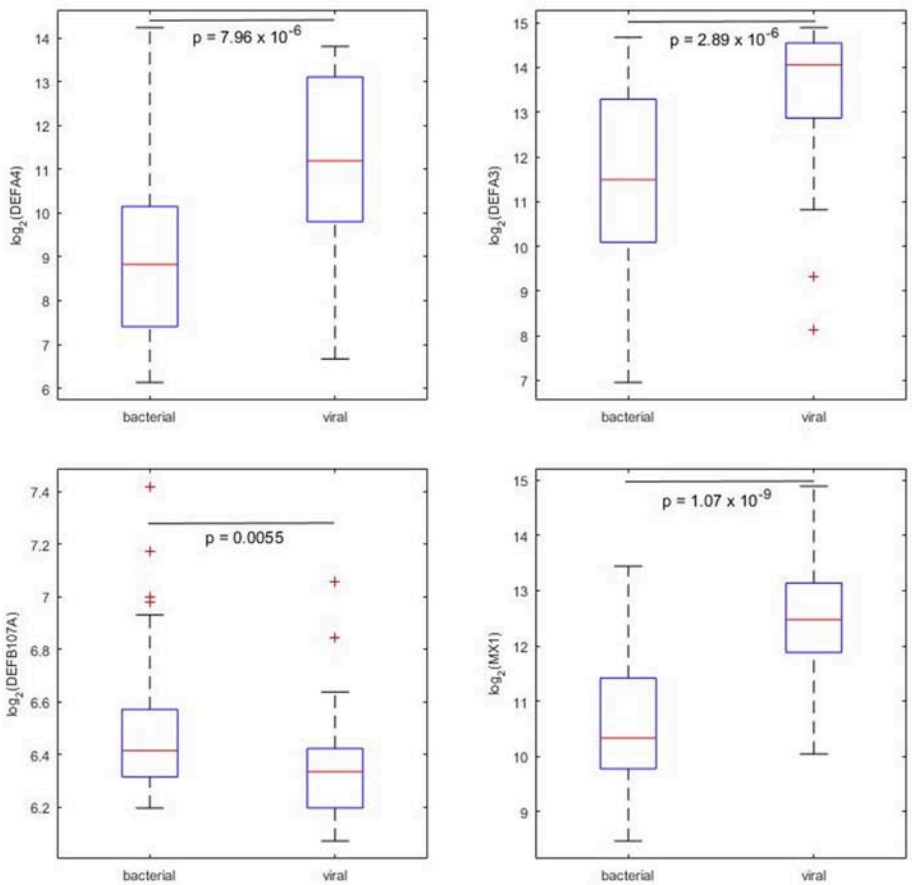


FIGURE 3 | Four example defensins and associated genes that showed significant differences between bacterial and viral pneumonia patients (illustrated by boxplots).

loads in each sample. To facilitate comparisons among samples, the bacterial/viral loads were normalized by sequencing depth (i.e., the total sequencing reads obtained for each sample).

Evaluating Correlations of Defensin Levels And Bacterial/Viral Loads

Spearman's rank correlation coefficient (Spearman, 1987) was then used to evaluate the associations between defensins and

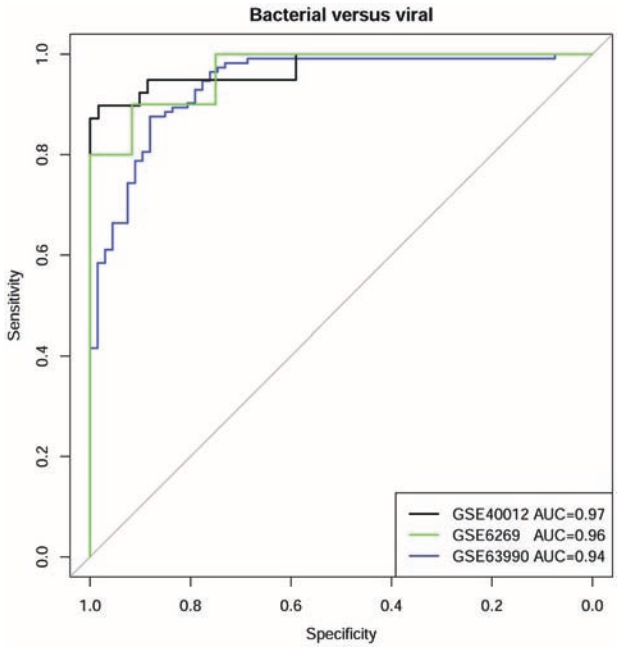


FIGURE 4 | ROC curves of defensins and associated genes for classifying bacterial and viral infections on three datasets.

viral/bacterial loads. Specifically, given the expression levels of a defensin at all the eight time points x_i where $i = 1, ..., 8$ and the normalized loads of a specific bacterial/viral species y_j where$j = 1, ..., 8$, ranks r^x and r^y were firstly obtained and then the correlation was calculated according to the following formula:

$$r_{xy} = \frac{\mathrm{cov}(r^x, r^y)}{\sigma_{r^x}\sigma_{r^y}} \quad (1)$$

Where $\mathrm{cov}(r^x, r^y)$ is the covariance of the rank variables and σ_{r^x} and σ_{r^y} are the standard deviations of the rank variables. For each pair of defensin and microbial species, the corresponding p-value was also calculated, which was further subject to multiple testing correction by the Benjamini and Hochberg method.

Validating the Diagnostic Value of Defensins On Independent Datasets

An independent cohort of 100 pneumonia patients (61 bacterial and 39 viral) were used to validate the diagnostic value of defensins and associated genes (NCBI Gene Expression Omnibus, access number: GSE40012) (Parnell et al., 2012). The whole blood gene expression profiles were quantified by Illumina HT-12 gene-expression beadarrays. Expression levels of defensins and associated genes were then extracted for clustering and classification analysis. For clustering analysis, t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) was first used to reduce the dimensionality of the data to two for visualization and then a clustering method based on searching density peaks (Rodriguez and Laio, 2014) was used to cluster the samples into two groups. For classification analysis, the popular random forest method (Breiman, 2001) was used to evaluate the diagnostic value via a leave-one-out cross-validation method. The diagnostic value of defensins and associated genes was further validated on two additional independent datasets. One dataset included 12 children's admitted to *Streptococcus pneumoniae* or Staphylococcus aureus infections and 10 children's admitted to viral infections by influenza viruses (NCBI Gene Expression Omnibus, access number: GSE6269) (Ramilo et al., 2007). The other dataset included 67 bacterial and 113 viral infections for adults (NCBI Gene Expression Omnibus, access number: GSE63990) (Tsalik et al., 2016).

RESULTS

Evident Associations of Different Defensins to the Bacterial And Viral Loads of H7N9 Pneumonia Patients

It is evident that influenza H7N9 virus demonstrated two peaks in the patient blood (from Day 6 to Day 12 and from Day 18 to Day 24), with days from Day 12 to Day 18 forming a valley (**Figure 1A**). However, at Day 18, a huge peak of *Acinetobacter baumannii* infection appeared which declined in the following days with small fluctuations (**Figure 1A**). The total of 30 defensins measured (4 α and 26 β defensins) were all expressed in at least one sample or more (**Table 1**). Most of the defensins except DEFA5, DEFB116, DEFB127, DEFB114, and DEFB108B did not show correlations to or only showed weak correlations to viral/bacterial loads in blood that were statistically not significant (**Figure 1B**). DEFA5 was positively correlated to the blood load of influenza A H7N9 virus ($r = 0.735$, $p < 0.05$, Spearman correlation), which also showed two peaks similar to those of the virus (**Figure 1A**). But DEFA5 did not show correlations to the bacterial load. Different from DEFA5,

DEFB116 and DEFB127 were positively correlated to the blood load of *Acinetobacter baumannii* ($r = 0.881$ and 0.810, $p < 0.05$), both of which showed two peaks with one consistent with the peak of *Acinetobacter baumannii* and another at Day 6 (**Figure 1A**). The peak at Day 6 may indicate latent bacterial infection that was undetectable in blood, suggesting potentially superior sensitivity of defensin-based diagnostics. DEFB114 and DEFB108B showed negative correlations with *Acinetobacter baumannii* ($r = -0.731$ and -0.786, $p < 0.05$, Spearman correlation, **Figures 1A,B**).

Diagnostic Values of Defensins On an Independent Pneumonia Cohort

On the independent validation dataset, we first extracted the expression profiles of defensins and associated genes and conducted t-SNE for visualization. It is obvious that bacterial and viral pneumonia patients separately formed clusters with a few exceptions (**Figure 2**, left). Clustering analysis grouped the patients into two classes, one of which corresponded to bacterial pneumonia and the other corresponded to viral pneumonia (**Figure 2**, middle). The accuracy of clustering analysis reached 82%, with 18 patients mis-clustered. Clustering based on the raw high-dimensional data resulted in similar results, suggesting

that bacterial and viral infections caused different responses for defensins and associated genes in blood. When switching the algorithms from unsupervised to supervised, high accuracy (93%), AUC (0.97), sensitivity (0.98), specificity (0.82), precision (0.90), and F1-score (0.94) were achieved by a random forest classifier with default parameters (**Figure 2**, right), suggesting the potential of defensin-based diagnostics to discriminate viral/bacterial infections.

Among the 87 defensins and associated genes that had expression values available, DEFA4 and DEFA3 were the most significantly differentially expressing defensins between bacterial and viral pneumonia patients. Both of these two defensins are alpha defensins and highly expressed in viral pneumonia patient blood (**Figure 3**, upper). The p-values tested by Wilcoxin rand-sum test were 7.96×10^{-6} and 2.89×10^{-6} for DEFA4 and DEFA3, respectively. DEFB107A was significantly highly expressed in bacterial pneumonia patient blood (**Figure 3**, lower left, $p = 0.0055$, Wilcoxin rand-sum test). MX1 is the most significant defensin-associated gene differentially expressed between bacterial and viral pneumonia (**Figure 3**, lower right, $p = 1.07 \times 10^{-9}$, Wilcoxin rand-sum test).

Evaluations on two additional datasets (GSE6269 and GSE63990) confirmed the diagnostic power of defensins and associated genes (**Figure 4**). On the dataset GSE6269, the accuracy can reach 95% while the AUC, sensitivity, specificity, precision, and F1-score are 0.96, 1, 0.9, 0.92, and 0.96, respectively. On the dataset GSE63990, similar performance was obtained, with accuracy 89%, AUC 0.94, sensitivity 0.84, specificity 0.93, precision 0.88 and F1-score 0.85.

DISCUSSION

Accurate discrimination of bacterial and viral infections has important clinical values and can inform clinicians to properly select therapies. Identifying biomarkers that can accurately classify bacterial infections from viral infections is thus of great importance. Blood-based assays including microarrays and next-generation sequencing provide a quite convenient method to quantify the expression levels of various genes, which form a rich resource for determination of biomarkers discriminating bacterial and viral infections. Multiple studies have been completed to seek such biomarkers from human blood gene expression profiles (Zaas et al., 2009; Parnell et al., 2012; Hu et al., 2013, 2015; Suarez et al., 2015). However, the values of defensins are often overlooked. Defensins, which are a major family of antimicrobial peptides expressed predominantly in neutrophils and epithelial cells and play important roles in innate immune defense against infectious pathogens, are hypothesized by us to action in distinct ways when combating against bacterial and viral infections, and thus we conducted this study.

We addressed the diagnostic values of defensins through two ways. Firstly, we checked the associations between human blood defensin mRNAs and the bacterial and viral loads through a continuous follow-up of a pneumonia patient caused by infection of influenza A H7N9 virus. This longitudinal study revealed that bacterial and viral loads were associated to beta and alpha defensins, respectively, among which several defensins showed impressing statistical significance. Secondly, we re-analyzed the diagnostic values of defensins on an independent dataset, which quantified blood gene expression profiles of 100 pneumonia patients including 61 bacterial and 39 viral infections. This lateral study demonstrated again the diagnostic power of defensins for discriminating bacterial and viral infections. Both studies remind that defensins and associated genes have great diagnostic potentials which deserve further investigation in the future. Although, the statistically significant defensins in these two studies did not overlap well, they could be caused or at least explained by the different study types and profiling techniques (microarray-based or NGS-based). Further studies were needed to exclude the technical interference and to include more biological variance.

We also compared the defensin-based biomarkers with published biomarker panels. We noticed that MX1 appeared multiple times across the studies, consistent with its great difference between bacterial and viral infections. Other defensins and associated genes are reported for the first time to have diagnostic power to discriminate bacterial from viral infections, and thus may provide new insights into the infection mechanisms and serve as important tools for clinical diagnosis. Because innate immunity is the first frontier of host to combat pathogens, the differences of defensins and associated genes during bacterial and viral infections may suggest that prominent patterns exist in host innate immune responses and defensins are valid representative molecules.

In summary, defensins not only are important molecules for hosts to combat infections, but also may provide promising biomarkers to indicate the types of infectious agents, which is expected to of significant clinical utility and needs further investigations.

AUTHOR CONTRIBUTIONS

XR, JY, and QJ designed the experiment. SZ and XR performed the experiment. XR wrote the manuscript with all the authors contributing to the writing.

FUNDING

This work was supported by the CAMS Innovation Fund for Medical Sciences (2017-I2M-3-017) and the National Key Research and Development Program (2016YFC1202404).

ACKNOWLEDGMENTS

The authors thank the members of Zhan group at Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences for their valuable discussion.

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2018). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 34:1440. doi: 10.1093/bioinformatics/btx773

Edelman, L. B., Toia, G., Geman, D., Zhang, W., and Price, N. D. (2009). Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases. *BMC Genomics* 10:583. doi: 10.1186/1471-2164-10-583

Ganz, T. (2003). Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol*, 3, 710–720 doi: 10.1038/nri1180

Hu, X., Yu, J., Crosby, S. D., and Storch, G. A. (2013). Gene expression profiles in febrile children with defined viral and bacterial infection. *Proc. Natl. Acad. Sci. U. S. A.* 110, 12792–12797. doi: 10.1073/pnas.1302968110

Hu, Y., Ren, X., Liu, Y., Yang, F., Liu, H., Cao, B., et al. (2015). Serial high-resolution analysis of blood virome and host cytokines expression profile of a patient with fatal H7N9 infection by massively parallel RNA sequencing. *Clin. Microbiol. Infect.* 21, 713.e1–4. doi: 10.1016/j.cmi.2015.03.006

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

Huang, Z. A., Chen, X., Zhu, Z., Liu, H., Yan, G. Y., You, Z. H., et al. (2017). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

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

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

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

McCullers, J. A. (2014). The co-pathogenesis of influenza viruses with bacteria in the lung. *Nat. Rev. Microbiol.* 12, 252–262. doi: 10.1038/nrmicro3231

Mejias, A., Dimo, B., Suarez, N. M., Garcia, C., Suarez-Arrabal, M. C., Jartti, T., et al. (2013). Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. *PLoS Med.* 10:e1001549. doi: 10.1371/journal.pmed.1001549

Mitra, S., Stärk, M., and Huson, D. H. (2011). Analysis of 16S rRNA environmental sequences using MEGAN. *BMC Genomics* 12 (Suppl. 3):S17. doi: 10.1186/1471-2164-12-S3-S17

Parnell, G. P., McLean, A. S., Booth, D. R., Armstrong, N. J., Nalos, M., Huang, S. J., et al. (2012). A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia. *Crit. Care* 16:R157. doi: 10.1186/cc11477

Peng, B., Zhu, D., Ander, B. P., Zhang, X., Xue, F., Sharp, F. R., et al. (2013). An integrative framework for Bayesian variable selection with informative priors for identifying genes and pathways. *PLoS ONE* 8:e67672. doi: 10.1371/journal.pone.0067672

Ramilo, O., Allman, W., Chung, W., Mejias, A., Ardura, M., Glaser, C., et al. (2007). Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* 109, 2066–2077. doi: 10.1182/blood-2006-02-002477

Rodriguez, A., and Laio, A. (2014). Clustering by fast search and find of density peaks. *Science* 344, 1492–6. doi: 10.1126/science.1242072

Spearman, C. (1987). The proof and measurement of association between two things. *Am J Psychol.* 100, 441–471. doi: 10.2307/1422689

Suarez, N. M., Bunsow, E., Falsey, A. R., Walsh, E. E., Mejias, A., and Ramilo, O. (2015). Superiority of transcriptional profiling over procalcitonin for distinguishing bacterial from viral lower respiratory tract infections in hospitalized adults. *J. Infect. Dis.* 212, 213–222. doi: 10.1093/infdis/jiv047

Sweeney, T. E., Wong, H. R., and Khatri, P. (2016). Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Sci. Transl. Med.* 8:346ra91. doi: 10.1126/scitranslmed.aaf7165

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Tsalik, E. L., Henao, R., Nichols, M., Burke, T., Ko, E. R., McClain, M. T., et al. (2016). Host gene expression classifiers diagnose acute respiratory illness etiology. *Sci. Transl. Med.* 8:322ra11. doi: 10.1126/scitranslmed.aad6873

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: http://www.jmlr.org/papers/v9/vandermaaten08a.html

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Ye, Y., Tsui, F. R., Wagner, M., Espino, J. U., and Li, Q. (2014). Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. *J. Am. Med. Inform. Assoc.* 21, 815–823. doi: 10.1136/amiajnl-2013-001934

Zaas, A. K., Chen, M., Varkey, J., Veldman, T., Hero, A. O., Lucas, J., et al. (2009). Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell. Host. Microbe* 6, 207–217. doi: 10.1016/j.chom.2009.07.006

Conflict of Interest Statement: 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.

Copyright © 2018 Zhou, Ren, Yang and Jin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 11 December 2018
doi: 10.3389/fmicb.2018.03076

Reconstruction and Analysis of a Genome-Scale Metabolic Model of *Ganoderma lucidum* for Improved Extracellular Polysaccharide Production

*Zhongbao Ma[1,2], Chao Ye[1,2], Weiwei Deng[2], Mengmeng Xu[1,2], Qiong Wang[1,2], Gaoqiang Liu[3], Feng Wang[4], Liming Liu[1,2], Zhenghong Xu[1,2], Guiyang Shi[1,2] and Zhongyang Ding[1,2]**

[1] Key Laboratory of Carbohydrate Chemistry and Biotechnology, Ministry of Education, School of Biotechnology, Jiangnan University, Wuxi, China, [2] National Engineering Laboratory for Cereal Fermentation Technology, Jiangnan University, Wuxi, China, [3] Key Laboratory of Cultivation and Protection for Non-Wood Forest Trees, Ministry of Education, College of Life Science and Technology, Central South University of Forestry and Technology, Changsha, China, [4] School of Food and Biological Engineering, Jiangsu University, Zhenjiang, China

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Aristóteles Góes-Neto,
Universidade Federal de Minas Gerais, Brazil
Mohammad Faiz Ahmad,
Jawaharlal Nehru University, India

***Correspondence:**
Zhongyang Ding
bioding@163.com

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

Received: *17 August 2018*
Accepted: *29 November 2018*
Published: *11 December 2018*

Citation:
Ma Z, Ye C, Deng W, Xu M, Wang Q, Liu G, Wang F, Liu L, Xu Z, Shi G and Ding Z (2018) Reconstruction and Analysis of a Genome-Scale Metabolic Model of Ganoderma lucidum for Improved Extracellular Polysaccharide Production. Front. Microbiol. 9:3076. doi: 10.3389/fmicb.2018.03076

In this study, we reconstructed for the first time a genome-scale metabolic model (GSMM) of *Ganoderma lucidum* strain CGMCC5.26, termed model *i*ZBM1060, containing 1060 genes, 1202 metabolites, and 1404 reactions. Important findings based on model *i*ZBM1060 and its predictions are as follows: (i) The extracellular polysaccharide (EPS) biosynthetic pathway was elucidated completely. (ii) A new fermentation strategy is proposed: addition of phenylalanine increased EPS production by 32.80% in simulations and by 38.00% in experiments. (iii) Eight genes for key enzymes were proposed for EPS overproduction. Model *i*ZBM1060 provides a useful platform for regulating EPS production in terms of system metabolic engineering for *G. lucidum*, as well as a guide for future metabolic pathway construction of other high value-added edible/ medicinal mushroom species.

Keywords: *Ganoderma lucidum*, extracellular polysaccharide, genome-scale metabolic model, biosynthetic pathway, phenylalanine, simulation

Abbreviations: Ac, acetate; Acal, acetaldehyde; Accoa, acetyl-CoA; Akg, alpha-2-oxoglutarate; Aol, D-arabinitol; Ara, arabinose; Ara1p, beta-arabinose-1-P; Cellob, cellobiose; Cit, citrate; D6pgc, 6-*P*-gluconate; D6pgl, 6-*O*-*P*-glucono-1,5-lactone; dTDP-Glc, dTDP-alpha-glucose; dTDP-Oglc, dTDP-4-oxo-6-deoxy-glucose; dTDP-Rmn, dTDP-4-dehydro-beta-rhamnose; E4p, erythrose-4-P; Eth, ethanol; Fdp, fructose-1,6-bisphosphate; Fru, fructose; Fuc, fucose; Fuc1p, fucose-1-P; Ful, fuculose; Fulp, fuculose-1-P; Fum, fumarate; G, glycerate; Ga3p, glyceraldehyde-3-P; Gal1p, alpha-galactose-1-P; GDP-Ddm, GDP-4-dehydro-6-deoxy-mannose; GDP-Fuc, GDP-fucose; GDP-Man, GDP-mannose; Gl, glycerol; Gl3p; sn-glycerol-3-P; Glac, galactose; Glx, glyoxylate; Glyal, glyceraldehyde; Glyn, glycerone; Icit, isocitrate; Kdr, 2-dehydro-3-deoxy-rhamnonate; Lact, lactose; Laol, L-arabinitol; Lxul, L-xylulose; Mal, malate; Man, mannose; Man1p, mannose-1-P; Man6p, mannose-6-P; Meli, melibiose; Mlt, maltose; Oaa, oxaloacetate; 13Pdg, 1,3-bisphosph-glycerate; 2pg, 2-*P*-glycerate; 3pg, 3-*P*-glycerate; Pep, phosphoenolpyruvate; Pyr, pyruvate; R5P, ribose-5-P; Rl, ribulose; Rl5p, ribulose-5-P; Rmf, rhamnofuranose; Rml, rhamnono-1,4-lactone; Rmn, rhamnose; Rmna, rhamnonate; S7p, sedoheptulose-7-P; Suc, sucrose; Succ, succinate; Succoa, succinyl-CoA; T3p2, glycerone-P; Tre, trehalose; UDP-Ara, UDP-arabinose; UDP-Ga, UDP-glucuronate; UDP-Glac, UDP-alpha-galactose; UDP-Glc, UDP-glucose; UDP-Xyl, UDP-xylose; Xol, xylitol; Xul, D-xylulose; Xul5p, xylulose-5-P; Xyl, xylose.

INTRODUCTION

Ganoderma lucidum (lingzhi or reishi mushroom) is a species well known for its edible and medicinal properties, and has a long history of use for prevention and treatment of various human diseases. Extracellular polysaccharides (EPSs) from *G. lucidum* comprise a structurally diverse group of macromolecules that display immunomodulatory, antitumor, and a wide range of other biological activities (Ferreira et al., 2015). Many studies have described enhancement of EPS production through optimization of medium and culture conditions in submerged fermentation (Tang and Zhong, 2003). However, EPS molecules, their structural features, and their biosynthetic pathways are all highly complex, and attempts to improve EPS production are often hampered by this complexity. There is an urgent need for more extensive, systematic knowledge of physiological features and metabolism of *G. lucidum* EPSs.

Genome-scale metabolic models (GSMMs), in which a systems biology approach is used to integrate genomic, transcriptomic, proteomic, and metabolomic data, are highly effective tools for metabolism research. GSMMs have been widely used for analysis of network properties, prediction of growth phenotypes, and interpretation of experimental data, particularly in *Escherichia coli* and *Saccharomyces cerevisiae* models (Kim et al., 2017).

There have been no reports to date of GSMMs for edible/medicinal mushroom species. The publication in 2012 of the whole genome sequence of *G. lucidum* strain CGMCC5.26 (Chen et al., 2012), and subsequent related reports, have made GSMM reconstruction feasible for this species. Such reconstruction will help clarify *G. lucidum* global metabolism, guide design of metabolic regulation strategies, and indicate useful research targets of "wet" experiments. Biosynthetic pathways of EPSs remain poorly known at this point because of our inadequate knowledge of related enzymes and their functions. Adequate knowledge will require gene cloning and genetic transformation studies (Wang et al., 2017).

We describe here reconstruction of the first GSMM of *G. lucidum*, model *i*ZBM1060, and its application to elucidate detailed physiological characteristics and production of EPSs in this species. The nucleoside sugar biosynthetic pathway of model *i*ZBM1060 was elucidated completely, the reactions of this pathway are summarized and illustrated, and related strategies for improving EPS production are proposed.

MATERIALS AND METHODS

Reconstruction and Refinement of *G. lucidum* GSMM

The availability of the whole genome sequence of *G. lucidum* allowed us to perform GSMM reconstruction according to a three-step general workflow scheme described previously (Thiele and Palsson, 2010).

(i) Sequenced *G. lucidum* genome data were downloaded from the UniProt database (UniProt, 2010). Genes were functionally annotated by two methods: (a) Thresholds of the bidirectional BLAST for a functional sequence were set to have *e*-value $< 1 \times 10^{-30}$, amino acid sequence identity $> 40\%$, and matching length $\geq 70\%$ of the query sequence (Liu et al., 2012). An original reactions list was produced by selecting GSMMs of *Aspergillus niger i*MA871(Andersen et al., 2008), *Mortierella alpina i*CY1106 (Ye et al., 2015), and *Aspergillus terreus i*JL1454 (Liu, et al., 2013) as template frameworks to map the assigned genes. (b) The KEGG Automatic Annotation Server (KAAS) (Moriya et al., 2007) was used for functional annotation of all amino acid query sequences.

(ii) A draft model was developed and used as a starting point for subsequent network refinements. Biochemical information was acquired from public databases [KEGG (Kanehisa et al., 2010), MetaCyc (Caspi et al., 2008), CELLO (Yu et al., 2004), and TCDB (Saier et al., 2006)], and manual revisions (deletion of error reactions, addition of organism-specific information, checking of mass-charge balance, filling of metabolic gaps) were conducted sequentially.

(iii) The COBRA Toolbox was used to simulate growth rate and product formation, and the model was validated by comparison of results with experimentally observed phenotypes (**Figure 1**).

Biomass Composition and Determination of Target Equation

The biomass components of *G. lucidum* are proteins, DNA, RNA, lipids, glucan, chitin and small molecules. Detailed information on biomass composition is summarized in **Supplementary Table S1**. A metabolic model (Andersen et al., 2008) were used as reference to calculate ATP required for cell growth and RNA: DNA ratio. Nucleotide and amino acid compositions were calculated based on *G. lucidum* genome (Chen et al., 2012). Detailed compositions of individual macromolecules were derived from published reports on *G. lucidum* (Mau et al., 2001; Stojkovic et al., 2014). A target equation of EPS production was determined based on mole percentages of monosaccharides in EPSs (Peng et al., 2015).

G. lucidum Strain and Culture Conditions

Ganoderma lucidum CGMCC5.26 was obtained from the China General Microbiological Culture Collection Center (Beijing) and maintained on potato dextrose agar slants at 4°C. The seed and fermentation medium [glucose 20 g/L, yeast nitrogen base without amino acids (YNB) 5 g/L, tryptone 5 g/L, KH_2PO_4 4.5 g/L, $MgSO_4{\cdot}7H_2O$ 2 g/L, initial pH 6.0] was kept at 30°C on a rotary shaker (150 rpm). The minimal growth medium for functional tests was composed of carbon source 20 g/L, nitrogen source 10 g/L, KH_2PO_4 4.5 g/L, $MgSO_4{\cdot}7\ H_2O$ 2 g/L, initial pH 6.0.

Determination of Biomass, Residual Sugar in Medium, and EPS

Mycelia were harvested by centrifugation (10,000 rpm) for 10 min. The precipitate was washed three times with distilled water, and dried at 60°C to constant weight. Dry weight (DW) was determined by gravimetric method. Amount of residual

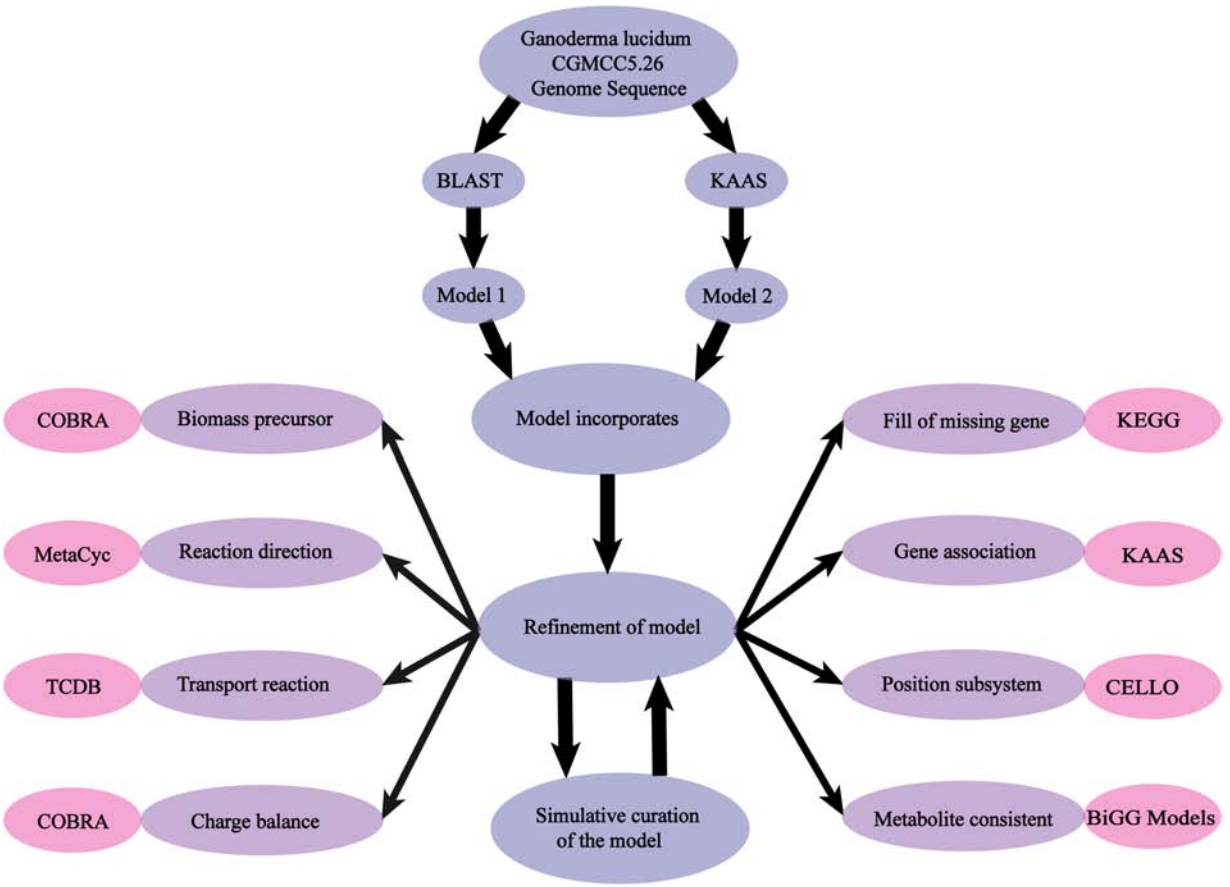


FIGURE 1 | Scheme of model reconstruction process used in this study. Blue: basic process of model reconstruction. Purple: major aspects of refinement of the model. Pink: usage of Database and Toolbox to refine the process.

sugar in medium was determined by 3, 5-dinitrosalicylic acid (DNS) method (Dubois et al., 1956).

For determination of EPS, centrifugal fluid as above was precipitated with adding 4 times of 95% (v/v) ethanol and left 8 h at 4°C to precipitate crude polysaccharides. Precipitate was collected by centrifugation (8,000 rpm) for 20 min, washed three times with 80% (v/v) ethanol, and dried at 60°C to remove residual ethanol. Total EPS content was assayed by phenol-sulfuric acid method (Dubois et al., 1951).

Simulation, Curation, and Analysis of Model *i*ZBM1060

To assess the ability of the reconstruction to accurately reflect metabolic processes of *G. lucidum*, we converted the reaction list to a standard SBML document that could be read by COBRA Toolbox (Schellenberger et al., 2011) and subjected to Flux balance analysis (FBA) (Lakshmanan et al., 2014). Flux ranges of reactions in the network were limited for simulations (Thiele and Palsson, 2010). Essential elements must be obtained from the environment through the exchange reaction (Wang et al., 2016). For growth simulation, the biomass equation in minimal medium (no amino acids) was set as the objective function. A complex fermentation medium (basic elements and 20 amino acids) was simulated for EPS production, and maximal uptake rate for each amino acid was set to 0.01 mmol/gDW/h (Ye et al., 2015). Essential genes were assessed by setting fluxes of reactions to zero, and simulating optimal growth rate with FBA. The criterion for an essential gene was that its deletion results in zero growth.

For identification of target genes, MOMA (Segre et al., 2002) framework was used for better prediction of flux distribution. The overexpression algorithm involved five steps (Boghigian et al., 2012): (i) EPS production flux was imposed on the reconstructed model. (ii) Flux for each reaction was calculated based on the fermentation medium. (iii) Amplification of flux was imposed on individual reactions with non-zero flux, to simulate the effect of gene overexpression. (iv) MOMA was performed to overcome the problem of overexpression. (v) An overexpressed target having higher EPS production and f_{PH} value > 1 was identified (Equation 1), f_{PH} being the product of the specific biomass overexpression and specific EPS overexpression rates.

$$f_{PH} = (f_{biomass})(f_{EPS}) = \left(\frac{V_{biomass,\,overexpression}}{V_{biomass,\,WT}}\right)\left(\frac{V_{EPS,\,overexpression}}{V_{EPS,\,WT}}\right) \quad (1)$$

RESULTS AND DISCUSSION

Reconstruction and Characteristics of Model *i*ZBM1060

The GSMM reconstruction was completed by automatic annotation and manual refinement, and a reaction list was obtained through KAAS and BLASTP. The final reconstructed GSMM of *G. lucidum*, termed model *i*ZBM1060, contained 1060 genes, 1202 metabolites, and 1404 reactions (**Supplementary Table S2**). The 1404 reactions in model *i*ZBM1060 were classified into 10 subsystems, according to the KEGG Pathway Database (**Figure 2A**). The largest subsystem (accounting for 21.97% of the 1404 reactions) was lipid metabolism (fatty acid biosynthesis; fatty acid degradation; glycerolipid, glycerophospholipid, sphingolipid, and steroid metabolism), followed by amino acid metabolism and carbohydrate metabolism. These three subsystems, combined, accounted for >50% of the 1404 reactions. There were a total of 1047 gene-associated reactions. In eight of the 10 subsystems, >80% of the reactions were associated with genes (the exceptions were lipid metabolism and transport reactions; **Figure 2B**).

Growth Verification and Simulation in Model *i*ZBM1060

Qualitative Verification and Analysis of Growth Phenotypes

The central metabolic pathway of *G. lucidum* carbon sources is shown schematically in **Figure 3**. Glucose, galactose, mannose, and fructose produce a corresponding phosphate monosaccharide through action of a kinase, and the monosaccharide then passes directly into the tricarboxylic acid (TCA) cycle, glyoxylate cycle, and pentose phosphate pathway (PPP). Xylose and arabinose are first phosphorylated by oxidation-reduction reaction, and then fructose-6-phosphate (fructose-6-P) is synthesized. In rhamnose and fucose metabolism, pyruvate and glycerone phosphate (respectively) are synthesized firstly.

The capability of *G. lucidum* to utilize 18 different carbon sources (13 saccharides, 3 alcohols, 2 carboxylic acids) for cell growth was predicted qualitatively by FBA. Each of the carbon sources was used as sole carbon source in minimal growth medium. Results were compared to experimental data, and the growth phenotype matching rate was 94.4%(**Table 1**). *G. lucidum* is able to utilize not only glucose, galactose, mannose, arabinose, xylose, rhamnose, fucose, and other monosaccharides, but also sucrose, maltose, lactose, and other disaccharides. FBA also predicted the capability to utilize various nitrogen sources (nitrate, urea, 20 amino acids) for cell growth. When results were compared to experimental data, the matching rate was 95.5% (**Table 2**).

Ganoderma lucidum grew successfully on 17 of the 18 carbon sources and 20 of the 22 nitrogen sources as above, indicating its broad substrate adaptability. There were no "fatal gaps" in model *i*ZBM1060, and it can therefore be used for predicting catabolic pathways of various carbon and nitrogen sources. Two of the apparently non-conforming sources (citrate and urea) can be attributed to unclear transport pathways and the absence of regulatory mechanisms in this stoichiometric model.

On the basis of carbon source metabolic pathways and the experimental results, we selected seven monosaccharides as single carbon sources for evaluation of effects of various carbon sources on biomass and EPS production (**Supplementary Figure S1** and **Figure 4A**). The consistency of results further supports the validity of the model.

Quantitative Verification

Fermentation data were used as constraints for simulation of cell growth, including specific growth rate and glucose uptake rate. Maximal specific growth rate was 0.076 h^{-1}, and corresponding sugar consumption rate was 0.506 mmol/gDW/h (**Figure 4B**). For simulation of cell growth in various media, the biomass equation was maximized in flux analysis. For glucose medium

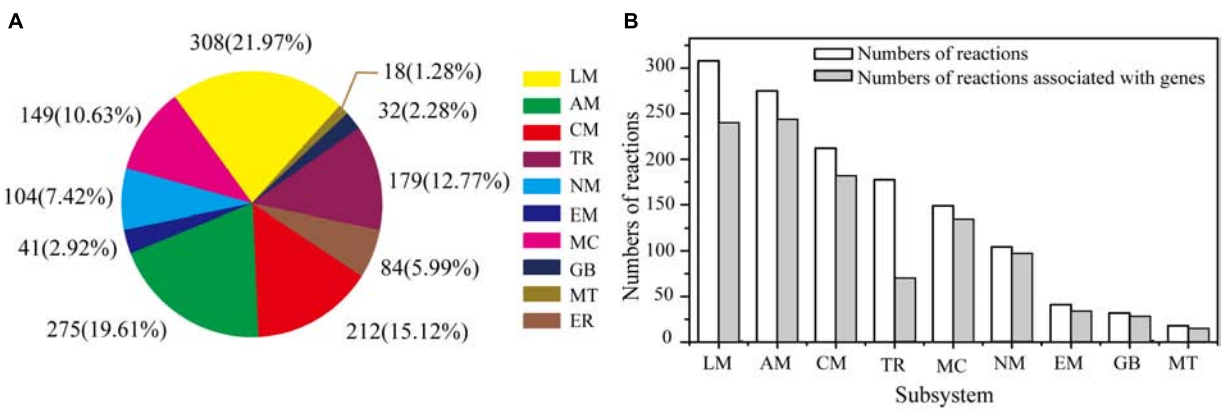


FIGURE 2 | Properties of model *i*ZBM1060. **(A)** Numbers of reactions in each subsystem, and corresponding percentages of total reactions. **(B)** Numbers of reactions associated with genes in each subsystem. LM, lipid metabolism; AM, amino acid metabolism; CM, carbohydrate metabolism; TR, transport reactions; MC, metabolism of cofactors and vitamins; NM, nucleotide metabolism; EM, energy metabolism; GB, glycan biosynthesis and metabolism; MT, metabolism of terpenoids and polyketides; ER, exchange reactions.

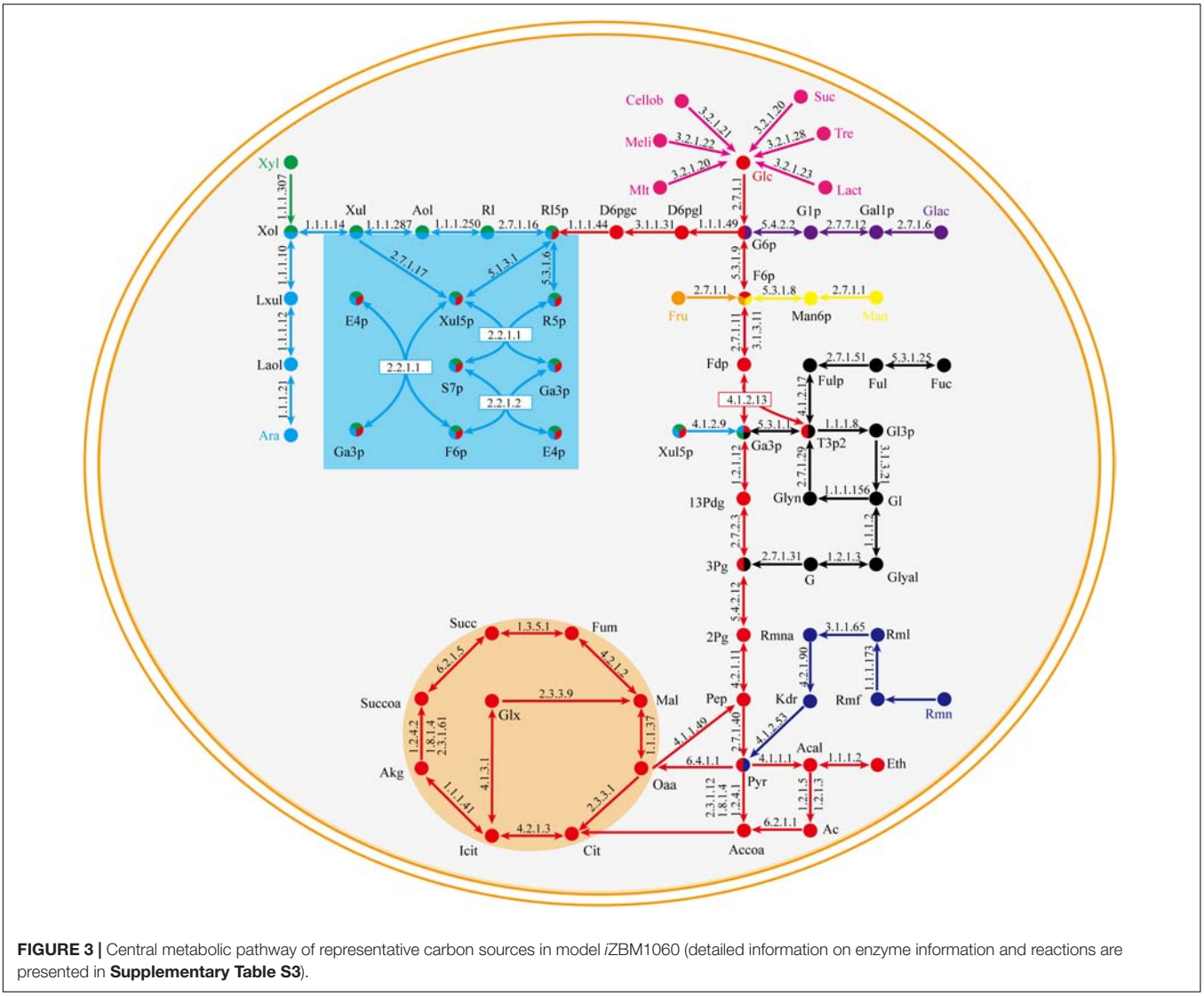


FIGURE 3 | Central metabolic pathway of representative carbon sources in model *i*ZBM1060 (detailed information on enzyme information and reactions are presented in **Supplementary Table S3**).

without production constraints, predicted cell growth rate was 0.077 h^{-1} – only 1.3% higher than experimental growth rate (0.076 h^{-1}).

Identification and Analysis of Essential Genes for Cell Growth

Consistency of growth rate *in silico* and *in vivo* indicated that model *i*ZBM1060 successfully reflected *G. lucidum* cellular metabolism. Essential genes for cell growth were predicted by single-gene deletion in COBRA Toolbox (MATLAB package) with two media (minimal growth medium, fermentation medium). Hundred and nineteen genes (11.23% of 1060 total genes) were predicted to be essential in minimal growth medium, and 88 genes (8.30% of total) were predicted to be essential in fermentation medium (**Figure 5A**). On minimal growth medium, >50% of the essential genes for growth were involved in either amino acid (32.77%) or carbohydrate metabolism (19.33%) (**Figure 5B**). In contrast, on fermentation medium, >90% of essential genes for growth were classified in 5 subsystems (amino acid metabolism, metabolism of cofactors and vitamins, nucleotide metabolism, carbohydrate metabolism, lipid metabolism) (**Figure 5C**), reflecting the important roles of these subsystems in cell growth (essential genes and simulation conditions are listed in **Supplementary Table S4**).

Nucleoside Sugar Biosynthetic Pathway in *G. lucidum*

Construction of Nucleoside Sugar Biosynthetic Pathway

The biosynthetic pathway of EPSs can be divided into three stages: (i) biosynthesis of nucleoside sugar precursors; (ii) assembly of repeating units; (iii) process of polymerization (Li et al., 2016). The monosaccharide composition of all EPSs includes glucose, galactose mannose, xylose, arabinose, fucose, and rhamnose. Typically, the proportion of glucose is high whereas that of fucose and rhamnose is low (Peng et al., 2015;

TABLE 1 | Growth phenotypic validation under a sole carbon source.

Substrate	Biomass		
Carbon source	***In vivo***	***In silico***	**Reference**
Glucose	+	+	Babitskaya et al., 2005; Wei et al., 2016
Sucrose	+	+	Babitskaya et al., 2005; Wei et al., 2016
Galactose	+	+	Babitskaya et al., 2005; Wei et al., 2016
Mannose	+	+	Babitskaya et al., 2005; Wei et al., 2016
Xylose	+	+	Babitskaya et al., 2005; Wei et al., 2016
Maltose	+	+	Babitskaya et al., 2005; Wei et al., 2016
Lactose	+	+	Babitskaya et al., 2005; Wei et al., 2016
Fructose	+	+	Babitskaya et al., 2005
Arabinose	+	+	Babitskaya et al., 2005
Mannitol	+	+	Babitskaya et al., 2005
Cellobiose	+	+	Babitskaya et al., 2005
Starch	+	+	Babitskaya et al., 2005
Fucose	+	+	Babitskaya et al., 2005
Rhamnose	+	+	This study
Inositol	−	−	This study
Ethanol	+	+	This study
Citrate	−	+	This study
Malate	+	+	This study

In vivo, experimental results. In silico, simulation results.

TABLE 2 | Growth phenotypic validation under a sole nitrogen source.

Substrate	Biomass		
Nitrogen source	***In vivo***	***In silico***	**Reference**
Urea	−	+	This study
NH4Cl	+	+	This study
L-Methionine	−	−	This study
L-Isoleucine	+	+	This study
L-Leucine	+	+	This study
L-Phenylalanine	+	+	This study
L-Proline	+	+	This study
L-Alanine	+	+	This study
L-Glutamate	+	+	This study
L-Glutamine	+	+	This study
L-Glycine	+	+	This study
L-Threonine	+	+	This study
L-Aspartate	+	+	This study
L-Asparagine	+	+	This study
L-Tryptophan	+	+	This study
L-Histidine	+	+	This study
L-Serine	+	+	This study
L-Tyrosine	+	+	This study
L-Valine	+	+	This study
L-Lysine	+	+	This study
L-Arginine	+	+	This study
L-Cysteine	+	+	This study

In vivo, experimental results. In silico, simulation results.

Wang et al., 2017). Monosaccharide heterogeneity is reflected in the complexity of EPS biosynthetic pathways.

Biosynthetic pathways of EPSs are poorly known because of our inadequate knowledge of related enzymes and their functions. On the basis of model *i*ZBM1060, we hereby propose a detailed nucleoside sugar biosynthetic pathway. Glucose, galactose, fucose, mannose, and arabinose reactions are catalyzed by monosaccharide kinase to produce corresponding phosphate monosaccharides, and UDP-glucose, UDP-galactose, GDP-mannose, and UDP-arabinose are then synthesized through action of pyrophosphorylase. E.g., hexokinase (GL26783-R1, GL20491-R1, and GL20491-R2) and mannose phosphomutase (GL20742-R1 and GL21817-R1) participate respectively in synthesis of mannose-6-P and mannose-1-P. GDP-mannose is then synthesized by action of GDP-mannose pyrophosphorylase (GL25424-R1).

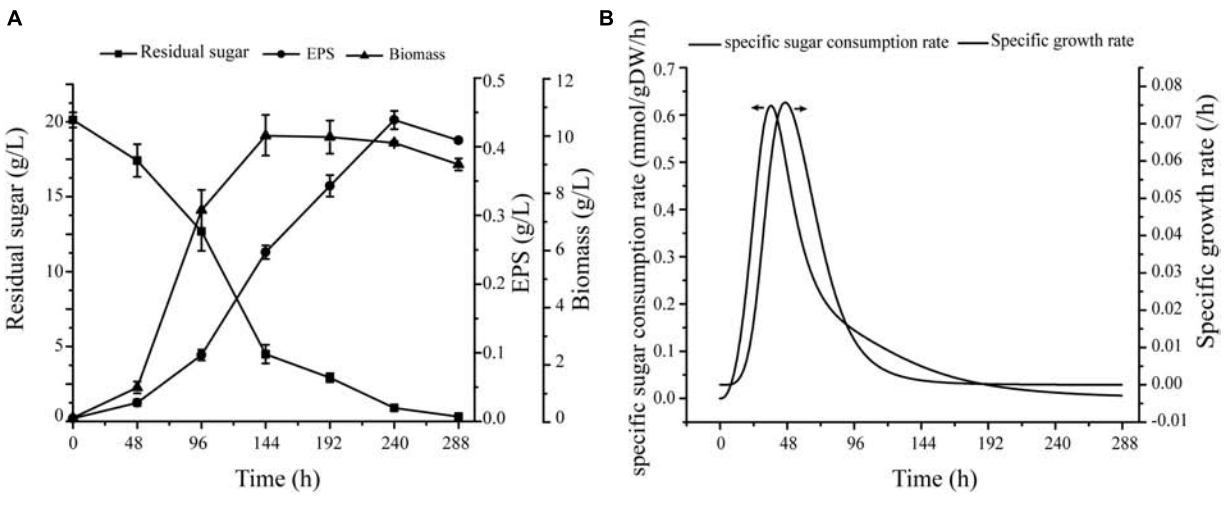


FIGURE 4 | Verification of specific growth rates in cells. **(A)** Under conditions of fermentation medium, cell growth, EPS production, and sugar consumption curve. **(B)** Nonlinear curve fit of growth curve and sugar consumption curve.

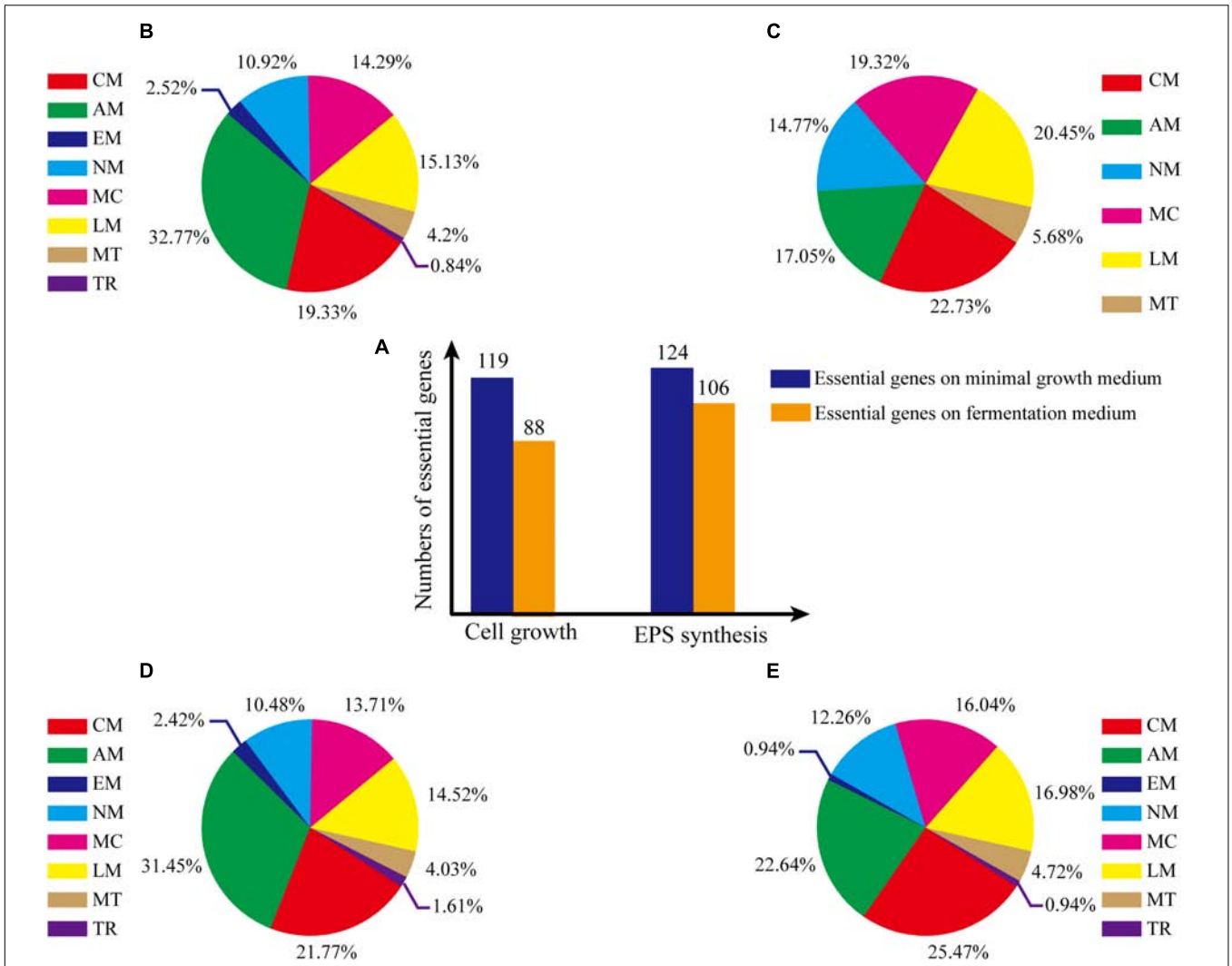


FIGURE 5 | Percentages of essential genes in each subsystem under various culture conditions. **(A)** Numbers of essential genes under various conditions. **(B)** Essential genes for cell growth, on minimal growth medium. **(C)** Essential genes for cell growth, on fermentation medium. **(D)** Essential genes for EPS production, on minimal growth medium. **(E)** Essential genes for EPS production, on fermentation medium. The abbreviations of each subsystem are the same with that in **Figure 2**.

Xylose enters the PPP to synthesize fructose-6-P, and then a nucleoside precursor. Fucose and rhamnose can also synthesize fructose-6-P via the gluconeogenesis pathway. Fructose-6-P has two pathways for synthesis of nucleoside sugar: (i) glucose-6-P isomerase (GL22245-R1) catalyzes conversion of fructose-6-P to glucose-6-P; (ii) fructose-6-P is converted to mannose-6-P by mannose-6-P isomerase (GL17878-R1 and GL22193-R1) and further synthesizes GDP-mannose and GDP-fucose (**Figure 6**).

Glucose, galactose, mannose, fucose, and arabinose are able to synthesize nucleoside sugars via short metabolic pathways. In contrast, xylose, fructose, and rhamnose cannot directly enter the nucleoside sugar biosynthetic pathway, and are therefore less ideal carbon sources. This concept is supported by our "wet" experimental results (**Supplementary Figure S1** and **Figure 4A**).

The proposed nucleoside sugar biosynthetic pathway involves 20 genes, 17 enzymes, and glucose as carbon source. Peng et al. (2015) observed activity of related enzymes in a biosynthetic pathway, indicating the accuracy of our reconstructed pathway (**Table 3**).

Identification and Analysis of Essential Genes for EPS Synthesis

Essential genes for EPS synthesis were predicted by single-gene deletion using COBRA Toolbox in two media. Prior to such prediction, biomass function must be constrained to ensure normal growth of cells. For minimal growth medium and fermentation medium, 124 genes (11.70% of total) and 106 genes (10.00%), respectively, were identified as essential for EPS synthesis (**Figure 5A**; essential genes and simulation conditions are listed in **Supplementary Table S4**).

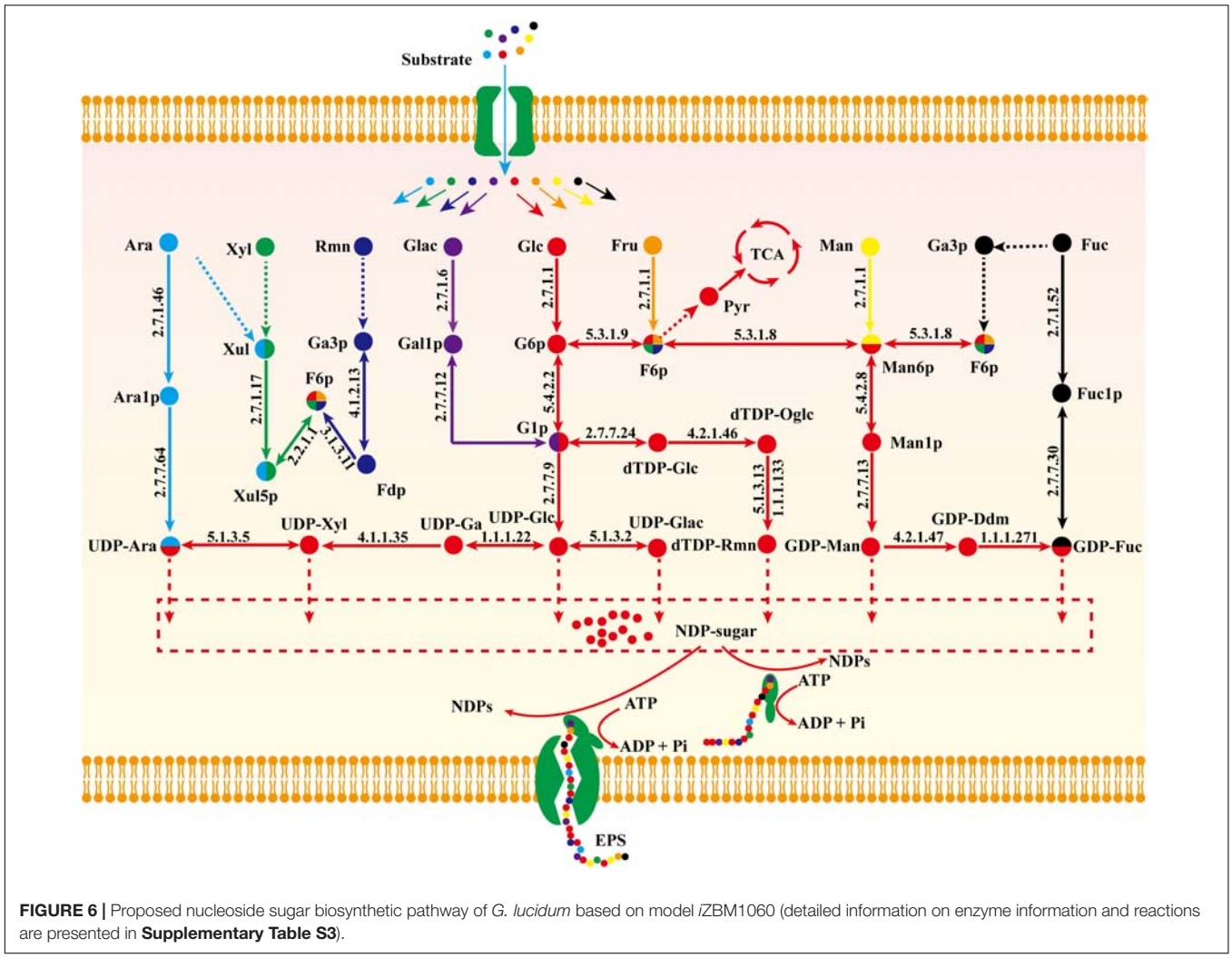


FIGURE 6 | Proposed nucleoside sugar biosynthetic pathway of *G. lucidum* based on model *i*ZBM1060 (detailed information on enzyme information and reactions are presented in **Supplementary Table S3**).

TABLE 3 | The reported enzymes of *G. lucidum* EPS biosynthetic pathway (Peng et al., 2015).

EC No.	Gene ID	Enzyme name	Reaction
5.3.1.9	GL22245-R1	Glucose-6-phosphate isomerase (GPI)	D-Glucose-6-phosphate <=> D-Fructose-6-phosphate
3.1.3.11	GL24591-R1 GL24591-R2	Fructose-1,6-bisphosphatase (FBP)	D-Fructose-1,6-bisphosphate - > D-Fructose-6-phosphate
5.3.1.8	GL17878-R1 GL22193-R1	Mannose-6-phosphate isomerase (PMI)	D-Mannose-6-phosphate <=> D-Fructose-6-phosphate
1.1.1.22	GL18437-R1	UDP-glucose-6-dehydrogenase (UGDH)	UDP-glucose <=> UDP-glucuronate
5.4.2.2	GL24280-R1	Glucose phosphomutase (PGM)	D-Glucose 1-phosphate <=> D-Glucose-6-phosphate
2.7.7.9	GL25739-R1	UTP-glucose pyrophosphorylase (UGP)	D-Glucose-1-phosphate - > UDP-glucose
2.7.7.24	Not annotated	dTDP-glucose synthase (RFFH)	D-Glucose-1-phosphate <=> dTDP-glucose
5.1.3.2	GL30389-R1 GL29575-R1	UDP-glucose 4-epimerase (GALE)	UDP-glucose <=> UD*P*-alpha-D-galactose
2.7.7.13	GL25424-R1	GDP-mannose pyrophosphorylase (GMP)	D-Mannose-1-phosphate - > GDP-mannose

For minimal growth medium, predicted essential genes for EPS synthesis were involved primarily in amino acid metabolism (31.45%) and carbohydrate metabolism (21.77%) (**Figure 5D**). For fermentation medium, predicted essential genes were involved in amino acid metabolism (22.64%), metabolism of cofactors and vitamins (16.04%), lipid metabolism (16.98%, and carbohydrate metabolism (25.47%) (total ∼80%; **Figure 5E**). These findings indicate that more carbon metabolism pathways are needed for EPS synthesis on minimal growth medium than on fermentation medium.

Comparative Genomics Analysis of Nucleoside Sugar Biosynthetic Pathway

To further elucidate EPS metabolic mechanisms and pathways, we performed comparative genomics analysis of *G. lucidum* EPSs and other important fungi.

TABLE 4 | Key enzymes of EPS biosynthesis in six well-studied mushroom species.

EC No.	*G. lucidum*	*A. cinnamomea*	*C. militaris*	*O. sinensis*	*F. velutipes*	*P. ostreatus*
2.7.1.1	3	2	4	4	2	2
5.3.1.9	1	1	1	1	1	1
5.3.1.8	2	2	1	1	1	1
5.4.2.8	2	0	0	0	0	0
2.7.7.13	1	0	0	0	0	0
4.2.1.47	1	1	0	0	1	1
1.1.1.271	1	1	0	0	1	1
5.4.2.2	1	2	1	1	3	2
2.7.7.9	1	2	1	1	1	1
5.1.3.2	2	1	1	1	1	1
1.1.1.22	1	1	1	1	1	1
4.1.1.35	1	1	0	0	1	1
5.1.3.5	0	0	0	0	0	0
2.7.7.24	0	0	0	0	0	0
4.2.1.46	0	0	1	2	0	0
5.1.3.13	0	0	0	0	0	0
1.1.1.133	0	0	0	0	0	0
2.7.1.6	1	0	1	1	1	1
2.7.7.12	0	0	1	1	1	1
1.1.1.307	1	0	0	0	0	0
1.1.1.9	1	1	2	0	1	1
1.1.1.14	1	1	2	3	3	2
4.1.2.13	2	1	3	2	2	2
3.1.3.11	2	1	1	1	1	1
2.7.1.17	2	2	1	1	2	2
2.2.1.1	0	1	1	1	1	1
2.7.1.11	1	1	1	1	1	1
4.1.2.9	1	1	0	0	1	1
2.7.1.46	1	0	0	0	0	0
2.7.7.64	1	0	0	0	0	0
2.7.1.52	1	0	0	0	0	0
2.7.7.30	1	0	0	0	0	0

Numbers in table columns are numbers of genes that were annotated.

Genomes of five related edible/ medicinal mushroom species (*Antrodia cinnamomea* (Riley et al., 2014), *Cordyceps militaris* (Zheng et al., 2011), *Ophiocordyceps sinensis* (Hu et al., 2013), *Flammulina velutipes* (Park et al., 2014), *Pleurotus ostreatus* (Riley et al., 2014) were annotated by KAAS. Metabolic enzymes related to EPS biosynthesis in these other species were compared with those in *G. lucidum* to clarify the characteristics of EPS biosynthesis. A total of 32 key enzymes were annotated. Numbers of key enzymes annotated were 25 for *G. lucidum*, 18 for *A. cinnamomea*, 17 for *C. militaris*, 16 for *O. sinensis*, 20 for *F. velutipes*, and 20 for *P. ostreatus* (**Table 4**). *G. lucidum* and the other five species had comprehensive gene annotations in the glycolysis pathway. All six species displayed gene deletion in the dTDP-rhamnose biosynthetic pathway. For example, dTDP-glucose synthase participated in synthesis of dTDP-glucose, and dTDP-4-dehydro-6-deoxy- D-glucose 3,5-epimerase and dTDP-4-dehydrorhamnose reductase participated in synthesis of dTDP-rhamnose. UDP-arabinose 4-epimerase, which catalyzed conversion of UDP-xylose to UDP-arabinose, was absent.

The above findings, taken together, indicate that the metabolic network of *G. lucidum* is more complex than other fungi, and allows synthesis of a greater variety of fungal nucleoside sugar precursors.

Optimization Strategies for Improving EPS Production *in silico*

On the basis of our model analysis, we propose two feasible optimization strategies for improvement of EPS production.

Biochemical Engineering Strategies

Effects of addition of amino acids on cell growth and EPS production were simulated, and both these parameters were

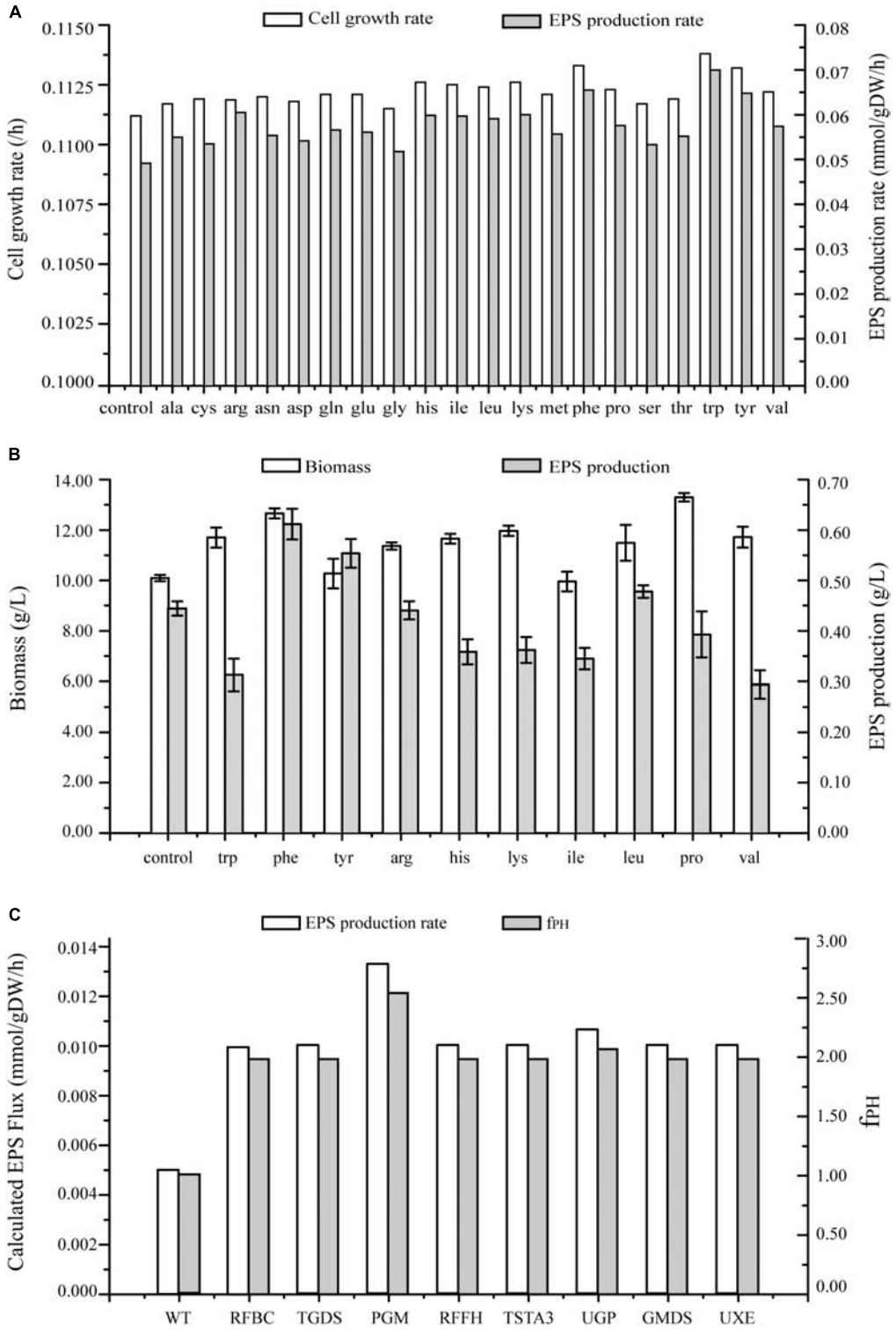


FIGURE 7 | Predicted EPS production rates in *G. lucidum* under various conditions. **(A)** Maximal growth rate and EPS production rate on minimal growth medium, and on minimal growth medium with one of 20 amino acids. **(B)** Effects on biomass and EPS production rate of separate addition of 10 amino acids. **(C)** Effects of single gene overexpression on EPS production rate and f_{PH} (see Simulation, Curation, and Analysis of Model iZBM1060). RFBC, dTDP-4-dehydro-6-deoxy-D-glucose 3,5-epimerase; TGDS, dTDP-glucose 4,6-hydro-lyase; PGM, glucose phosphomutase; RFFH, dTDP-glucose pyrophosphorylase; TSTA3, GDP-fucose synthase; UGP, UDP-glucose pyrophosphorylase; GMDS, GDP-mannose 4,6-hydro-lyase; UXE, UDP-arabinose 4-epimerase.

found to be enhanced (**Figure 7A**). EPS production was increased > 15% by addition of 10 separate amino acids in simulation results. Especially, the increase was greatest for tryptophan (41.85%), followed by phenylalanine (32.80%) and tyrosine (31.39%).

The wet experiment results indicated that EPS production was greatly increased by addition of phenylalanine (38.00%) or tyrosine (25.00%) and no enhancement occurred when tryptophan or the other amino acids were applied in EPS production (**Figure 7B**). Addition of phenylalanine is clearly effective in enhancing EPS production. Polysaccharide yields can be further improved by adjustment of quantity and timing of amino acid addition in future studies.

Genetic Engineering Strategies

In several recent studies, expression levels of EPS biosynthetic genes have been manipulated in order to increase EPS production. However, less is known regarding overexpression of target genes for this purpose. On the basis of GSMM, we simulated gene overexpression to guide metabolic engineering for enhancement of EPS production, using MOMA to reevaluate the fluxes and obtain an overexpression algorithm. Eight key enzymes were identified as potential targets for EPS production; i.e., overexpression of PGM gene (EC: 5.4.2.2, GL24280-R1), UGP gene (EC: 2.7.7.9, GL25739-R1), TSTA3 gene (EC: 1.1.1.271, GL21002-R1), GMDS gene (EC: 4.2.1.47, GL20928-R1), UXE gene (EC: 5.1.3.5), RFBC gene (EC: 5.1.3.13), TGDS gene (EC: 4.2.1.46), and RFFH gene (EC: 2.7.7.24) notably enhanced EPS production (**Figure 7C** and **Supplementary Table S5**).

PGM catalyzed conversion of glucose-6-P to glucose-1-P; each of these compounds is an important intermediate in EPS biosynthetic pathway. Overexpression of PGM gene increased EPS production from 0.005 mmol/gDW/h in wild-type (WT) to 0.0133 mmol/gDW/h. Thus, increased PGM transcription level was directly correlated with increased EPS production. PGM was also implicated as the key enzyme for EPS biosynthesis in a previous study: maximal EPS production in a PGM gene-overexpressing strain was 1.76 g/L – 44.3% higher than in WT (Xu et al., 2015). Overexpression of UGP gene *in silico* caused an increase of EPS production rate to 0.0106 mmol/gDW/h. UGP is directly involved in synthesis of UDP-glucose, and EPSs contain a high proportion of glucose. UDP-glucose plays a key role in EPS production as a synthetic precursor. Li et al. (2015) also demonstrated an effect of UGP on EPS synthesis.

UXE catalyzes interconversion of two EPS synthesis precursors: UDP-arabinose and UDP-xylose. When glucose is used as carbon source, UXE plays an essential role in production of UDP-arabinose. TSTA3 and GMDS are involved in synthesis of GDP-fucose. RFBC, TGDS, and RFFH participate in synthesis of dTDP-rhamnose.

Results of the analysis described above indicate that EPS production can be effectively improved by overexpression of genes for eight key enzymes. Previous studies have demonstrated the usefulness of PGM and UGP genes in this regard. Future studies will focus on overexpression of the other six genes for improvement of EPS production.

CONCLUSION

A GSMM for *Ganoderma lucidum* (lingzhi mushroom) is presented here for the first time. The GSMM (termed model *i*ZBM1060) is focused on EPSs, and contains 1404 reactions, 1202 metabolites, and 1060 genes. The model was validated and shown to accurately simulate cell growth and EPS production under various conditions. The nucleoside sugar (EPS precursor) biosynthetic pathway in the model was elucidated completely. Essential genes for cell growth and EPS synthesis, and genes for eight key EPS production enzymes, were analyzed. Two strategies for improvement of EPS production, based on model *i*ZBM1060, were proposed: (i) addition of phenylalanine; (ii) overexpression of the eight key enzyme genes. PGM and UGP genes have previously been shown to be useful targets for enhancement of EPS production, and future studies will focus on overexpression of the other six genes for this purpose. Model iZBM1060 provides a useful platform for regulating EPS production in terms of system metabolic engineering for *G. lucidum*, as well as a guide for future metabolic pathway construction of other high value-added edible/ medicinal mushroom species.

AUTHOR CONTRIBUTIONS

ZM and ZD designed the experiments. ZM, WD, QW, and MX performed the experiments. ZM, CY, GL, FW, GS, ZD, LL, and ZX conceived the project, analyzed the data, and wrote the paper.

FUNDING

This study was supported by the National Natural Science Foundation of China (No. 31271918), the China Postdoctoral Science Foundation (2015M571691) and the Qing Lan Project.

ACKNOWLEDGMENTS

The authors are grateful to S. Anderson for English editing of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.03076/full#supplementary-material

REFERENCES

Andersen, M. R., Nielsen, M. L., and Nielsen, J. (2008). Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Mol. Syst. Biol.* 4:178. doi: 10.1038/msb.2008.12

Babitskaya, V. G., Shcherba, V. V., Puchkova, T. A., and Smirnov, D. A. (2005). Polysaccharides of *Ganoderma lucidum* factors affecting their production. *Appl. Biochem. Microbiol.* 41, 169–173. doi: 10.1007/s10438-005-0029-1

Boghigian, B. A., Armando, J., Salas, D., and Pfeifer, B. A. (2012). Computational identification of gene over-expression targets for metabolic engineering of taxadiene production. *Appl. Microbiol. Biotechnol.* 93, 2063–2073. doi: 10.1007/s00253-011-3725-1

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., et al. (2008). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome databases. *Nucleic Acids Res.* 36, D623–D631. doi: 10.1093/nar/gkm900

Chen, S., Xu, J., Liu, C., Zhu, Y., Nelson, D. R., Zhou, S., et al. (2012). Genome sequence of the model medicinal mushroom *Ganoderma lucidum*. *Nat. Commun.* 3:913. doi: 10.1038/ncomms1923

Dubois, M., Gilles, K. A., Hamilton, J. K., and Rebers, P. A. (1951). A colorimetric method for the determination of sugars. *Nature* 168, 167–167. doi: 10.1038/168167a0

Dubois, M., Gilles, K. A., Hamilton, J. K., Rebers, P. A., and Smith, F. (1956). Colorimetric method for determination of sugars and related substances. *Anal. Chem.* 28, 350–356. doi: 10.1021/ac60111a017

Ferreira, I. C., Heleno, S. A., Reis, F. S., Stojkovic, D., Queiroz, M. J., Vasconcelos, M. H., et al. (2015). Chemical features of Ganoderma polysaccharides with antioxidant, antitumor and antimicrobial activities. *Phytochemistry* 114, 38–55. doi: 10.1016/j.phytochem.2014.10.011

Hu, X., Zhang, Y. J., Xiao, G. H., Zheng, P., Xia, Y. L., Zhang, X. Y., et al. (2013). Genome survey uncovers the secrets of sex and lifestyle in caterpillar fungus. *Chin. Sci. Bull.* 58, 2846–2854. doi: 10.1007/s11434-013-5929-5

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360. doi: 10.1093/nar/gkp896

Kim, W. J., Kim, H. U., and Lee, S. Y. (2017). Current state and applications of microbial genome-scale metabolic models. *Curr. Opin. Syst. Biol.* 2, 10–18. doi: 10.1016/j.coisb.2017.03.001

Lakshmanan, M., Koh, G., Chung, B. K., and Lee, D. Y. (2014). Software applications for flux balance analysis. *Brief. Bioinform.* 15, 108–122. doi: 10.1093/bib/bbs069

Li, H. J., Zhang, D. H., Yue, T. H., Jiang, L. X., Yu, X., Zhao, P., et al. (2016). Improved polysaccharide production in a submerged culture of *Ganoderma lucidum* by the heterologous expression of *Vitreoscilla* hemoglobin gene. *J. Biotechnol.* 217, 132–137. doi: 10.1016/j.jbiotec.2015.11.011

Li, M. J., Chen, T. X., Gao, T., Miao, Z. G., Jiang, A. L., Shi, L., et al. (2015). UDP-glucose pyrophosphorylase influences polysaccharide synthesis, cell wall components, and hyphal branching in *Ganoderma lucidum* via regulation of the balance between glucose-1-phosphate and UDP-glucose. *Fungal Genet. Biol.* 82, 251–263. doi: 10.1016/j.fgb.2015.07.012

Liu, J., Gao, Q., Xu, N., and Liu, L. M. (2013). Genome-scale reconstruction and in silico analysis of Aspergillus terreus metabolism. *Mol. Biosyst.* 9, 1939–1948. doi: 10.1039/c3mb70090a

Liu, T., Zou, W., Liu, L. M., and Chen, J. (2012). A constraint-based model of *Scheffersomyces stipitis* for improved ethanol production. *Biotechnol. Biofuels* 5:11. doi: 10.1186/1754-6834-5-72

Mau, J. L., Lin, H. C., and Chen, C. C. (2001). Non-volatile components of several medicinal mushrooms. *Food Res. Int.* 34:521. doi: 10.1016/s0963-9969(01)00067-9

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321

Park, Y. J., Baek, J. H., Lee, S., Kim, C., Rhee, H., Kim, H., et al. (2014). Whole genome and global gene expression analyses of the model mushroom *Flammulina velutipes* reveal a high capacity for lignocellulose degradation. *PLoS One* 9:e93560. doi: 10.1371/journal.pone.0093560

Peng, L., Qiao, S. K., Xu, Z. H., Guan, F., Ding, Z. Y., Gu, Z. H., et al. (2015). Effects of culture conditions on monosaccharide composition of *Ganoderma lucidum* exopolysaccharide and on activities of related enzymes. *Carbohyd. Polym.* 133, 104–109. doi: 10.1016/j.carbpol.2015.07.014

Riley, R., Salamov, A. A., Brown, D. W., Nagy, L. G., Floudas, D., Held, B. W., et al. (2014). Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. *Proc. Natl. Acad. Sci. U.S.A.* 111, 9923–9928. doi: 10.1073/pnas.1400592111

Saier, M. H. Jr., Tran, C. V., and Barabote, R. D. (2006). TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res.* 34, D181–D186. doi: 10.1093/nar/gkj001

Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models the COBRA Toolbox v2. 0. *Nat. Protoc.* 6, 1290–1307. doi: 10.1038/nprot.2011.308

Segre, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15112–15117. doi: 10.1073/pnas.232349399

Stojkovic, D. S., Barros, L., Calhelha, R. C., Glamoclija, J., Ciric, A., van Griensven, L. J., et al. (2014). A detailed comparative study between chemical and bioactive properties of *Ganoderma lucidum* from different origins. *Int. J. Food Sci. Nutr.* 65, 42–47. doi: 10.3109/09637486.2013.832173

Tang, Y. J., and Zhong, J. J. (2003). Role of oxygen supply in submerged fermentation of *Ganoderma lucidum* for production of Ganoderma polysaccharide and ganoderic acid. *Enzyme Microb. Technol.* 32, 478–484. doi: 10.1016/s0141-0229(02)00338-1

Thiele, I., and Palsson, B. O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121. doi: 10.1038/nprot.2009.203

UniProt, C. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142–D148. doi: 10.1093/nar/gkp846

Wang, B., Ning, Q. J., Hao, T., Yu, A. L., and Sun, J. S. (2016). Reconstruction and analysis of a genome-scale metabolic model for *Eriocheir sinensis* eyestalks. *Mol. Biosyst.* 12, 246–252. doi: 10.1039/c5mb00571j

Wang, Q., Wang, F., Xu, Z. H., and Ding, Z. Y. (2017). Bioactive mushroom polysaccharides a review on monosaccharide composition, biosynthesis and regulation. *Molecules* 22:E955. doi: 10.3390/molecules22060955

Wei, Z. H., Liu, L., Guo, X. F., Li, Y. J., Hou, B. C., Fan, Q. L., et al. (2016). Sucrose fed-batch strategy enhanced biomass, polysaccharide, and ganoderic acids production in fermentation of *Ganoderma lucidum* 5.26. *Bioprocess. Biosyst. Eng.* 39, 37–44. doi: 10.1007/s00449-015-1480-x

Xu, J. W., Ji, S. L., Li, H. J., Zhou, J. S., Duan, Y. Q., Dang, L. Z., et al. (2015). Increased polysaccharide production and biosynthetic gene expressions in a submerged culture of *Ganoderma lucidum* by the overexpression of the homologous alpha-phosphoglucomutase gene. *Bioprocess. Biosyst. Eng.* 38, 399–405. doi: 10.1007/s00449-014-1279-1

Ye, C., Xu, N., Chen, H., Chen, Y. Q., Chen, W., and Liu, L. M. (2015). Reconstruction and analysis of a genome-scale metabolic model of the oleaginous fungus *Mortierella alpina*. *BMC Genomics* 16:799. doi: 10.1186/s12864-015-2042-y

Yu, C. S., Lin, C. J., and Hwang, J. K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci.* 13, 1402–1406. doi: 10.1110/ps.03479604

Zheng, P., Xia, Y. L., Xiao, G. H., Xiong, C. H., Hu, X., Zhang, S. W., et al. (2011). Genome sequence of the insect pathogenic fungus *Cordyceps militaris*, a valued traditional chinese medicine. *Genome Biol.* 12:21. doi: 10.1186/gb-2011-12-11-r116

Conflict of Interest Statement: 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.

Copyright © 2018 Ma, Ye, Deng, Xu, Wang, Liu, Wang, Liu, Xu, Shi and Ding. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 19 December 2018
doi: 10.3389/fmicb.2018.03112

A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data

Jian Xiao[1,2], *Li Chen*[3]*, *Yue Yu*[1], *Xianyang Zhang*[4] and *Jun Chen*[1]*

[1] Division of Biomedical Statistics and Informatics, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, United States, [2] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China, [3] Department of Health Outcomes Research and Policy, Harrison School of Pharmacy, Auburn University, Auburn, AL, United States, [4] Department of Statistics, Texas A&M University, College Station, TX, United States

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Jonathan Badger,
National Cancer Institute (NCI), United States
Adina Howe,
Iowa State University, United States

***Correspondence:**
Li Chen
li.chen@auburn.edu
Jun Chen
chen.jun2@mayo.edu

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

Received: *31 August 2018*
Accepted: *03 December 2018*
Published: *19 December 2018*

Citation:
Xiao J, Chen L, Yu Y, Zhang X and Chen J (2018) A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data. Front. Microbiol. 9:3112. doi: 10.3389/fmicb.2018.03112

Fueled by technological advancement, there has been a surge of human microbiome studies surveying the microbial communities associated with the human body and their links with health and disease. As a complement to the human genome, the human microbiome holds great potential for precision medicine. Efficient predictive models based on microbiome data could be potentially used in various clinical applications such as disease diagnosis, patient stratification and drug response prediction. One important characteristic of the microbial community data is the phylogenetic tree that relates all the microbial taxa based on their evolutionary history. The phylogenetic tree is an informative prior for more efficient prediction since the microbial community changes are usually not randomly distributed on the tree but tend to occur in clades at varying phylogenetic depths (*clustered signal*). Although community-wide changes are possible for some conditions, it is also likely that the community changes are only associated with a small subset of "marker" taxa (*sparse signal*). Unfortunately, predictive models of microbial community data taking into account both the sparsity and the tree structure remain under-developed. In this paper, we propose a predictive framework to exploit *sparse* and *clustered* microbiome signals using a phylogeny-regularized sparse regression model. Our approach is motivated by evolutionary theory, where a natural correlation structure among microbial taxa exists according to the phylogenetic relationship. A novel phylogeny-based smoothness penalty is proposed to smooth the coefficients of the microbial taxa with respect to the phylogenetic tree. Using simulated and real datasets, we show that our method achieves better prediction performance than competing sparse regression methods for sparse and clustered microbiome signals.

Keywords: microbiome, phylogenetic tree, sparse generalized linear model, predictive model, statistical modeling, high-dimenisonal statistics

1. INTRODUCTION

The human microbial community (a.k.a., microbiota) is the collection of microorganisms associated with the human body. These microorganisms, their genomes, and the environment they reside in are collectively known as the human "microbiome." The human microbiome plays a critical role in health and disease (Cho and Blaser, 2012). For instance, the human gut microbiome aids the digestive system with inaccessible nutrients, synthesizes beneficial nutrients and protects us against pathogens. An abnormal microbiome has been implicated in many human diseases including various cancer types (Ahn et al., 2013; Bultman, 2014; Walther-Antonio et al., 2016; Peters et al., 2017). Dysbiosis of the microbiome has been observed in obesity, type II diabetes, rheumatoid arthritis and multiple sclerosis (Turnbaugh et al., 2009; Kinross et al., 2011; Honda and Littman, 2012; Pflughoeft and Versalovic, 2012; Qin et al., 2012; Chen et al., 2016; Jangi et al., 2016). Therefore, the human microbiome holds great potential for various clinical applications such as disease diagnosis, patient stratification and drug response prediction. Building up an efficient microbiome-based predictor could thus empower microbiome-based precision medicine (Kashyap et al., 2017).

Advances in low-cost, high-throughput DNA sequencing technologies such as Illumina Solexa sequencing has enabled researchers to study the microbiome composition by directly sequencing the microbial DNA. Two main approaches have been employed to sequence the microbiome: gene-targeted sequencing and shotgun metagenomic sequencing (Kuczynski et al., 2011). Compared to the shotgun metagenomic sequencing, where all microbial DNA is sequenced, the gene-targeted approach only sequences a "fingerprint" region of a "molecular clock" gene such as the 16S rRNA gene in the bacteria. Although the shotgun metagenomic sequencing provides more biological information, the targeted approach is still the dominant approach for large-scale microbiome studies due to its lower cost and high scalability (McDonald et al., 2018). In the targeted sequencing, standard practices involve clustering the sequencing reads into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) based on their sequence similarities (Schloss et al., 2009; Caporaso et al., 2010, 2012; Chen et al., 2013b, 2017; Edgar, 2013; Rideout et al., 2014; Callahan et al., 2016; Amir et al., 2017). A taxonomic lineage is further assigned to each OTU/ASV by comparing their sequence to existing 16S rRNA gene databases. Finally, a phylogenetic tree, which characterizes the evolutionary relationships among OTUs/ASVs, is constructed based on their sequence divergences (Price et al., 2010). For shotgun metagenomic sequencing, a phylogenetic tree can also be constructed based on the reference genomes of the detected species (Kembel et al., 2011). As a result, a typical microbiome sequencing study is usually summarized as a table of the read counts of the detected OTUs/ASVs/Species, together with a phylogenetic tree, reflecting the community structure and composition of the studied microbiome. For simplicity, hereafter, we use the term "OTU" to stand for the basic taxonomic units (e.g., OTU, ASV, species, taxa) from any sequencing experiment/bioinformatics pipeline. Compared to other types of omics sequencing data, one important characteristic of microbiome sequencing data (microbial community data) is the phylogenetic tree that relates all the OTUs. The phylogenetic tree provides prior knowledge about how the OTUs are evolutionarily related. Related OTUs, which usually share similar biological functions, are more likely to be simultaneously associated with the outcome, forming "clustered signals" at varying phylogenetic depths (Garcia et al., 2014; Martiny et al., 2015). Therefore, the phylogeny creates linkages among OTUs and induces a grouping structure, allowing more efficient linkage between the OTUs and the phenotype. As the microbial community data moves into even higher resolutions such as strain-level resolution (Mallick et al., 2017; Edgar, 2018), the phylogenetic relationship becomes even more important for OTU data analysis. Clearly, it is not sensible to treat OTUs with only 1% sequence divergence in the same way as the OTUs with more than 10% sequence divergence. Indeed, incorporating the tree structure has proven to make the analyses more efficient and robust for various statistical tasks ranging from ordination to microbiome-wide multiple testing (Purdom, 2011; Chen et al., 2012, 2013a; Evans and Matsen, 2012; Wang and Zhao, 2017; Xiao et al., 2017).

One important task for microbiome analysis is to predict the phenotype/outcome (either quantitative or qualitative) based on the features of the underlying microbial community (relative abundances of the OTUs and their phylogeny). This process is also known as predictive modeling or supervised learning in machine learning literature, where we try to derive some function from the training data that can be used to predict the outcome of future data, and to learn which features (i.e., OTUs) are predictive of the outcome. For clinical applications, the outcome includes disease state, treatment response, and drug toxicity. To enable prediction based on microbial community data, general-purpose predictive methods have been applied (Knights et al., 2011; Statnikov et al., 2013; Pasolli et al., 2016). These methods include classical machine learning methods (e.g., Random Forest and Support Vector Machine) and modern regression methods for high-dimensional data [e.g., Lasso (Tibshirani, 1996), MCP (Zhang, 2010), and Elastic Net (Zou and Hastie, 2005)], focusing on modeling the nonlinear relationship between the outcome and the microbiome as well as selecting the most predictive OTUs for better interpretation. However, these methods do not fully exploit the information in the microbiome data, particularly the phylogenetic relationship among OTUs. The phylogenetic tree is an informative prior since the microbial community changes are usually not randomly distributed but tend to occur in clades at varying phylogenetic depths (*clustered signal*). In other words, the phylogenetic structure offers a biologically motivated grouping structure, through which we can aggregate sparse OTU data to enrich signals and achieve better predictive performance. The objective of the proposed study is thus to provide a data-adaptive approach to use the tree structure when constructing the predictive model, i.e., let the data determine how much phylogenetic information and what level of phylogenetic depth we should use to achieve optimal performance. The inputs of our method are the OTU count table, the phylogenetic tree of the OTUs and the outcome measurements, and the outputs are

the selected OTUs and the predictive function based on their abundances.

Many previous attempts have been made to incorporate the tree information into prediction, particularly in the regression framework (Tanaseichuk et al., 2014; Chen et al., 2015; Ning and Beiko, 2015; Wang and Zhao, 2017; Randolph et al., 2018; Xiao et al., 2018). These methods are advantageous over previous methods by taking into account the tree. However, they still have many limitations. For example, some methods do not perform variable selection in model building (Wang and Zhao, 2017; Randolph et al., 2018; Xiao et al., 2018), and hence their prediction performance is subpar for sparse-signal scenarios (i.e., only a subset of OTUs are associated with outcome). For methods that perform variable weighting or selection (Tanaseichuk et al., 2014; Ning and Beiko, 2015), they usually rely solely on the tree topology. The branch lengths, which provide more detailed evolutionary history, are usually ignored. Therefore, there is still a need to develop prediction methods for sparse clustered signals while exploiting the full information of the phylogenetic tree, which consists of both the tree topology and branch lengths.

Previously, we developed *glmgraph* (Chen et al., 2015), a graph-regularized sparse regression model for structured genomic data. In the *glmgraph* framework, besides a sparsity penalty, a graph Laplacian-based structure penalty (Laplacian penalty) was imposed to smooth the coefficients with respect to the graph structure. It also encourages structurally related predictors to be selected simultaneously (Huang et al., 2011). In principle, a graph Laplacian can be constructed based on the pair-wise distances between OTUs with respect to the phylogenetic tree. However, the Laplacian penalty has two major drawbacks for microbiome applications. First, the Laplacian-induced smoothing/grouping effects are susceptible to the interference by a large number of distantly related OTUs since the graph is fully connected. It is well- known that distantly related OTUs have very different biological characteristics, and thus their contribution to the smoothing should be minimized. Second, the smoothing effects induced by the Laplacian penalty is completely driven by the external graph structure. This is in stark contrast to the l_2 penalty-induced smoothing effects (Zou and Hastie, 2005; Huang et al., 2016), which are mainly driven by the internal correlation structure in the data. In case of a misspecified tree, the Laplacian penalty cannot reduce to the l_2 penalty. Therefore, it does not possess the data-driven smoothing property, which has been shown to be important to improve prediction performance under certain scenarios (Waldron et al., 2011).

In this work, in parallel to our previous prediction method for "dense and clustered" microbiome signals (Xiao et al., 2018), we develop a phylogeny-regularized sparse regression model for "sparse and clustered" microbiome signals. The proposed method uses a novel phylogeny-based smoothness penalty, which is defined based on the inverse of the phylogeny-induced correlation matrix. The new penalty addresses the two major drawbacks of the Laplacian penalty: it encourages local smoothing, i.e., smoothing effects from more immediate neighbors, as well as enjoys the data-driven smoothing property if the tree is misspecified. In summary, the sparse nature of the distribution of OTUs in complex microbiome data can be better captured by our model because it provides a data-adaptive way to group the OTUs according to their phylogeny as well as to select the most predictive OTUs, which leads to improved prediction and interpretation.

2. METHODS

2.1. A Phylogeny-Induced Correlation Structure Among OTUs

We first introduce a phylogeny-induced correlation structure, on which our phylogeny-based smoothness penalty will be defined. Suppose we have p OTUs on a phylogenetic tree, following the evolutionary model proposed in Martins and Hansen (1997), the correlation of the traits between OTU i and j can be modeled as

$$c_{ij}(\alpha) = e^{-2\alpha d_{ij}},\ i, j = 1, \ldots, p, \tag{1}$$

where d_{ij} is the patristic distance between OTU i and j (i.e., the length of the shortest path linking the two OTUs on the tree) and the parameter $\alpha \in (0, \infty)$ characterizes the evolutionary rate. When $\alpha = 0$, $c_{ij} = 1\ \forall i, j$, indicating all the traits are the same and there is no evolution. When $\alpha \to \infty$, $c_{ij} = 0\ \forall i \neq j$, indicating that the traits evolve independently. The parameter α is also related to the phylogenetic depth of trait conservation (Martiny et al., 2015), with a smaller α value indicating a greater phylogenetic depth at which the trait is conserved (i.e., a large clade of OTUs share the trait). In other words, the parameter α has a (soft) grouping effect and groups the OTUs at various phylogenetic depths. Compared to the taxonomic grouping, where the OTUs are grouped at a specific taxonomic level, such phylogeny-based grouping not only achieves more resolutions, but also circumvents the difficulty of the uncertainty in taxonomy assignments. Therefore, in the context of predictive modeling, the parameter α can be treated as a tuning parameter, which allows us to explore different phylogenetic depths to optimize prediction. Also to be noted, the pairwise distance d_{ij} can be simply the genetic distance based on pairwise comparison of the DNA sequences without the need for explicit tree construction.

2.2. Phylogeny-Regularized Sparse Generalized Linear Model

To account for the high dimensionality and the phylogenetic tree structure in microbiome-based prediction, we introduce a phylogeny-regularized sparse generalized linear model. We assume that there are n samples with the abundances of p OTUs being profiled. For the ith sample, let y_i denote the outcome variable, which can be binary or continuous, and $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ denote the normalized and properly transformed abundance vector of the p OTUs. We further assume the data have been standardized ($\sum_i x_{ij} = 0, \sum_i x_{ij}^2 = n$). The goal is to predict y_i based on $\boldsymbol{x}_i$. We will use a generalized linear model

$$g(E(y_i)) = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ and $g(.)$ is a link function (identity and logit link for continuous and

binary outcome, respectively). Since $p > n$, we need to make some sparsity assumption in order for the model to be estimable. Additional assumption will be imposed on the structural relationship among the model parameters to make the estimation more efficient. To this end, we propose the following penalized log-likelihood to estimate the regression coefficients:

$$pl(\beta_0, \boldsymbol{\beta}; \lambda_1, \lambda_2) = \frac{1}{n}\sum_{i=1}^{n}\{-l(\beta_0, \boldsymbol{\beta}; y_i, \boldsymbol{x}_i)\} + p^{sp}_{\lambda_1}(\boldsymbol{\beta}) + p^{sm}_{\lambda_2}(\boldsymbol{\beta}), \tag{2}$$

where

$$l(\beta_0, \boldsymbol{\beta}; y_i, \boldsymbol{x}_i) = \begin{cases} -(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2/2 & \text{linear regression,} \\ y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) - \log(1 + e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}) & \text{logistic regression.} \end{cases}$$

The penalized likelihood estimate can be obtained by solving the optimization problem

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta_0, \boldsymbol{\beta}} pl(\beta_0, \boldsymbol{\beta}; \lambda_1, \lambda_2). \tag{3}$$

The two penalty terms in Equation (2) play distinct roles. $p^{sp}_{\lambda_1}(\boldsymbol{\beta})$ is the sparsity penalty, which induces a sparse solution and has been demonstrated to improve both the prediction performance and model interpretability (Tibshirani, 1996) in the high-dimensional setting. $p^{sm}_{\lambda_2}(\boldsymbol{\beta})$ is the smoothness penalty, which encourages smoothness of the estimated coefficients with respect to the phylogenetic tree (i.e., encourage similar coefficients for clustered OTUs at a certain phylogenetic depth).

For the sparsity penalty $p^{sp}_{\lambda_1}(\boldsymbol{\beta})$, we choose to use MCP (Minimax Concave Penalty) (Zhang, 2010):

$$p^{sp}_{\lambda_1}(\boldsymbol{\beta}) = \sum_{j=1}^{p}\rho(|\beta_j|; \lambda_1, \gamma), \quad \rho(t; \lambda_1, \gamma) = \lambda_1 \int_0^{|t|}(1 - x/(\gamma\lambda_1)_+ dx, \tag{4}$$

where $\lambda_1 \geq 0$ is the tuning parameter, $(.)_+$ indicates the nonnegative part and the parameter γ ($1 \leq \gamma \leq +\infty$) controls the degree of concavity. Larger values of γ make ρ less concave. By varying the value of γ from 1 to $+\infty$, the MCP provides a continuum of penalties with the hard-threshold penalty as $\gamma \to 1$ and the convex l_1 penalty at $\gamma = +\infty$. In practice, γ is usually fixed to a reasonable value without the need for further tuning. An important advantage of the MCP over the l_1 penalty is that it leads to a nearly unbiased estimator and achieves selection consistency under weaker conditions. More detailed discussions of MCP could be found in Zhang (2010).

Our major contribution is the design of a novel structure-based smoothness penalty $p^{sm}_{\lambda_2}(\boldsymbol{\beta})$ to achieve efficient phylogeny-based smoothing. One common approach to accommodate structure/graph information in sparse regression model is through the use of a graph Laplacian penalty $p^{sm}_{\lambda_2}(\boldsymbol{\beta}) = \lambda_2\boldsymbol{\beta}^T L\boldsymbol{\beta}$, where the Laplacian matrix L is defined based on the connectivity, or adjacency among predictors. The penalized likelihood estimator resulted from the combination of the MCP and Laplacian penalty, termed as *Sparse Laplacian Shrinkage* (SLS) estimator, has been shown to have nice properties such as selection consistency and generalized grouping (Huang et al., 2011). For microbiome applications, a graph Laplacian for microbiome data can be defined using the phylogeny-induced correlation (Equation 1) as the adjacency measure. However, we found that this approach did not always achieve better prediction performance than the procedure without the Laplacian penalty. The subpar performance is partly due to the interference by a large number of distantly related OTUs since the phylogeny-induced graph is fully connected. To achieve better prediction performance, it is important to reduce the contribution of smoothing effects from the large number of distantly related OTUs. Although this can be achieved by sparsifying L, in practice, the degree of sparsity to achieve optimal prediction depends on the data and it is difficult to set a universal degree of sparsity for all applications. To overcome the limitation of the graph Laplacian approach, we propose to use an alternative smoothness penalty

$$p^{sm}_{\lambda_2}(\boldsymbol{\beta}) = \lambda_2\boldsymbol{\beta}^T C^{-1}(\alpha)\boldsymbol{\beta}, \tag{5}$$

where $C(\alpha) = (c_{ij}(\alpha))_{p\times p}$ is the phylogeny-induced correlation structure defined in the previous section. The inverse correlation matrix $\Omega \triangleq C^{-1}$ also implies a graph structure among predictors but encourages more local smoothing, that is, the coefficient smoothing is mainly contributed by its immediate neighbors. To demonstrate a stronger local smoothing effect by Ω than L, we plot Ω_{ij}, L_{ij}, the elements of the Ω and L, against the pairwise patristic distances between OTUs (**Figure 1**). As the pairwise distance increases, Ω_{ij} approaches zero quickly while L_{ij} does not decrease as fast. Since $|\Omega_{ij}|, |L_{ij}|$ determine the contribution of the smoothing effect of OTU i to OTU j, a faster rate to zero suggests a stronger local smoothing effect.

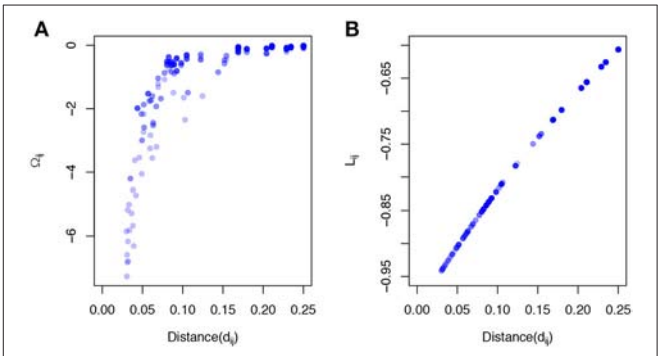


FIGURE 1 | Local smoothing effects of the proposed smoothness penalty. The data was generated based on a simulated phylogenetic tree ($p = 200$,"rcoal" from R "ape" package). The correlation $C(\alpha)$ was calculated based on the pairwise patristic distances with $\alpha = 2$. **(A)** The elements of inverse correlation matrix (Ω_{ij}) are plotted against pairwise patristic distances (d_{ij}). **(B)** The elements of Laplacian matrix (L_{ij}) are plotted against pairwise patristic distances (d_{ij}).

In the phylogeny-regularized sparse generalized linear model, we have three parameters λ_1, λ_2 and α, which need to be tuned in the training step for optimal prediction performance. These three parameters, respectively control the model sparsity (i.e., how many OTUs are predictive of the outcome), the phylogeny-based smoothing effects (i.e., how much smoothing effects should be induced by the tree), and the phylogenetic depth of the signal (i.e., what level of clustering is needed to achieve better prediction). With the inverse correlation matrix-based smoothness penalty, we call the resulting penalized likelihood estimator *Sparse Inverse Correlation Shrinkage* (SICS) estimator. The proposed approach also has a Bayesian interpretation: it assumes that the coefficient $\boldsymbol{\beta}$ has a prior multivariate normal component with the covariance matrix τC and the penalized likelihood estimate can be viewed as the MAP (maximum a posteriori) estimate from a Bayesian perspective.

2.3. Connection With Existing Methods

The proposed smoothness penalty $\boldsymbol{\beta}^T\Omega\boldsymbol{\beta}$, the graph Laplacian penalty $\boldsymbol{\beta}^T L\boldsymbol{\beta}$ and the l_2 penalty $\boldsymbol{\beta}^T\boldsymbol{\beta}$ are all special cases of a general class of quadratic penalties $\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}$, where Σ is a positive semi-definite matrix. When $\alpha \to \infty$, the proposed penalty becomes l_2 penalty and the SICS estimator is reduced to the Mnet estimator (Huang et al., 2016). It is well-known that l_2 penalty induces a grouping effect based on the correlation structure in the data (data-driven smoothing). As α decreases, the phylogeny-driven smoothing will take control (prior-driven smoothing). Thus, α also provides some tradeoff between data-driven and prior-driven smoothing (Theorem 1). To better understand the behavior of the proposed smoothness penalty, we rewrite it as

$$\boldsymbol{\beta}^T\Omega\boldsymbol{\beta} = \sum_{i=1}^{p}(\Omega_{ii} - \sum_{j=1,j\neq i}^{p}|\Omega_{ij}|)\beta_i^2 + \sum_{1\le j<k\le p}|\Omega_{jk}|(\beta_j - s_{jk}\beta_k)^2 \tag{6}$$

where $s_{jk} = \text{sgn}(-\Omega_{jk})$ is the sign of $-\Omega_{jk}$. Note that the second part has the same form as the Laplacian penalty (Huang et al., 2011). Thus, the proposed smoothness penalty is a combination of a weighted l_2 penalty (first part) and a Laplacian penalty (second part) with the adjacency coefficients $-\Omega_{ij}$. For the phylogeny-induced correlation structure, all the off-diagonal elements Ω_{ij} are negative and the magnitude controls the prior-driven smoothing effect. The weighted l_2 penalty, on the other hand, offers the data-driven smoothing effect. In contrast, the Laplacian penalty cannot reduce to the l_2 penalty and does not have the data-driven smoothing effect.

Since the proposed smoothness penalty has a weighted l_2 component, some degree of shrinkage in the coefficient estimate is expected (Zou and Hastie, 2005). For orthogonal designs, rescaling could remove the bias due to l_2 shrinkage without significantly increasing the variance. However, we find that, for more general designs, rescaling could instead increase the variance of the SICS estimator and decrease the prediction performance. Therefore, we will not rescale the coefficients in the implementation.

2.4. Some Theoretical Properties

We further investigate the smoothing effect and grouping property of the proposed SICS estimator. Previously, Li and Li (2008) derived the smoothing effect and grouping property for the penalty combining l_1 and Laplacian penalty, and Huang et al. (2016) demonstrated a similar property for the Mnet estimator. Here, we demonstrate such property for our SICS estimator under a linear regression model and a simple graph design. The proof of the theorem can be found in the **Supplementary File**.

Without loss of generality, we assume that the whole graph (as characterized by Ω) corresponding to the index set $\{1,\ldots,p\}$ is divided into disjoint cliques $V_1,\ldots,V_J$. We further assume that the patristic distances between OTUs are the same in each clique so that the phylogeny-induced correlation coefficient c_{ij} are the same. Thus, Ω has a special block-diagonal structure: $\Omega = \text{diag}(\Omega_1,\ldots,\Omega_J)$ with $\Omega_g=(\Omega_{g,lm})_{v_g\times v_g}$, where $v_g=|V_g|$ for $g=1,\ldots,J$, $\Omega_{g,ll} = \kappa_g(v_g-1)\Omega_g^0$ for $\Omega_g, \kappa_g>0, l=1,\ldots,v_g$ and $\Omega_{g,lm} = -\Omega_g^0$ for $1 \le l,m \le v_j, l \neq m$. Also, denote $\rho_{jk} = n^{-1}\sum_{i=1}^n x_{ij}x_{ik}$ (data-induced correlation between OTU i and OTU j). For the SICS estimator based on this inverse correlation matrix Ω, we have the following smoothing and grouping property:

Theorem 1. *Denote* $t = 2\lambda_2\kappa_g(v_g-1)\Omega_g^0$ *and*

$$\xi = \begin{cases} \max\{2\gamma(\gamma t-1)^{-1}, (\gamma t+1)(t(\gamma t-1))^{-1}, t^{-1}\}, & \text{if } \gamma t > 1, \\ t^{-1}, & \text{if } \gamma t \le 1. \end{cases}$$

Then for $j,k \in V_g$ *and* $g \in \{1,\ldots,J\}$, *we have*

$$|\hat{\beta}_j(\alpha,\lambda_1,\lambda_2) - \hat{\beta}_k(\alpha,\lambda_1,\lambda_2)| \le \frac{\xi||y||_1}{\sqrt{n}}\sqrt{2(1-\rho_{jk})}.$$

Especially, if $\rho_{jk}=0$, *we have* $|\hat{\beta}_j(\alpha,\lambda_1,\lambda_2) - \hat{\beta}_k(\alpha,\lambda_1,\lambda_2)| \le \frac{\sqrt{2}\xi||y||_1}{\sqrt{n}}$.

Based on Theorem 1, both the prior-induced correlation c_{jk} (which in turn determines Ω_g^0 and ξ) and the data-induced correlation ρ_{jk} contribute to the smoothing effect. With the tuning parameter α, c_{jk} can vary from 0 to 1 (equivalently, Ω_g^0 varies from 0 to ∞). We can thus increase and decrease the prior-driven smoothing by varying α. The optimal level of prior-driven smoothing effect can be tuned based on the data.

2.5. Model Estimation and Computational Complexity

Since the proposed penalty is convex with respect to $\boldsymbol{\beta}$, coordinate descent algorithm, which is developed for sparse regression model with convex and non-convex sparsity penalties (Friedman et al., 2010; Breheny and Huang, 2011) can be readily extended to our case. For the linear regression model, we have a closed-form solution for each coordinate update. For the logistic regression model, we solve a series of structure-regularized sparse linear regression model at each iteratively reweighed least squares step. Coordinate descent continues until a certain convergence criterion is reached. More details could

be found in Chen et al. (2015). We implemented the method in the R package *SICS* (https://github.com/lichen-lab/SICS), which depends on our previously developed *glmgraph* R package (Chen et al., 2015).

The computation complexity of the proposed method consists of two parts: coordinate descent and matrix inversion. For each coordinate descent loop, it requires $O(n + p)$ arithmetic operations, and a full cycle through the p OTUs requires $O(np + p^2)$ operations. Assume the number of iterations to reach convergence is c_1 and the number of tuning parameter combinations is c_2. The overall complexity for the coordinate descent algorithm is thus $O(c_1c_2(np + p^2))$. In addition, taking inverse of the correlation matrix typically has a computational complexity of $O(p^3)$ (some algorithm may reduce it, but could not bring down to $O(p^2)$). A total of $O(c_3p^3)$ is required to perform matrix inversion, where c_3 is the number of grid points for the tuning parameter α. Therefore, the total computational complexity for SICS is $O(c_1c_2(np + p^2) + c_3p^3)$. Usually, c_1, c_2, c_3 are treated as fixed, so the computational complexity for SICS is $O(np + p^3)$. Thus it is highly scalable with the sample size but not with the number of OTUs. Since we usually perform OTU filtering before running the algorithm, it is computationally efficient for typical microbiome datasets with $p < 1000$.

3. SIMULATION STUDIES

3.1. Simulation Strategy

We performed extensive simulations to evaluate the prediction performance of SICS for both continuous and binary outcome. For the continuous outcome, we simulated 100 samples in the training set and 200 samples in the testing set. For the binary outcome, we simulated an equal number of 50 samples for both case and control groups in the training set, and an equal number of 100 samples in case and control groups in the testing set. We used a Dirichlet-multinomial distribution with parameters estimated from a real microbiome data to simulate OTU counts and generated the outcome based on the abundances of the outcome-associated OTUs. We investigated the effect of the informativeness of the phylogenetic tree and the level of signal strength on the prediction performance. The simulation studies were aimed to reveal the scenarios in which our model performed favorably and also to test whether our model was robust when the phylogenetic tree was not informative or misspecified.

3.1.1. Simulating OTU Abundance Data

We included 200 OTUs in the simulation. The OTU counts were generated using a Dirichlet-multinomial distribution with the parameter values (dispersion, mean proportions) estimated based on a real dataset from the human upper respiratory tract microbiome (Charlson et al., 2010). Only the count data from the 200 most abundant OTUs were used in the parameter estimation. Accordingly, the phylogenetic tree was trimmed to contain the 200 OTUs. For each sample, the total read count was sampled from a negative binomial distribution with mean 5,000 and dispersion 25, reflecting a typical sequencing depth for a targeted sequencing experiment. The OTU counts were normalized into OTU proportions by dividing the total read counts.

3.1.2. Selecting Outcome-Associated OTUs

We simulated both phylogeny-informative and non-informative scenarios to study the performance of the proposed method with respect to the informativeness of the phylogenetic tree. In the phylogeny-informative scenarios, we selected outcome-associated OTUs ("aOTUs") from an OTU cluster and let their effects in the same direction. In the phylogeny-non-informative scenarios, we either randomly selected OTUs or let the effects of the aOTUs in a cluster have opposite effects, which violates the assumption that closely related aOTUs should have similar effects. To construct OTU clusters, we partitioned the 200 OTUs into 20 clusters using the partitioning-around-medoids (PAM) algorithm based on their patristic distances. The simulation strategy was illustrated in **Figure 2** and the detailed settings for four scenarios were presented below,

- S1: The phylogenetic tree was informative. One cluster with 12 aOTUs formed an outcome-associated cluster ("aCluster"). In the aCluster, the aOTUs had the same effect size and the effect direction was also the same.
- S2: The phylogenetic tree was informative. On top of S1, we varied the effect size of each aOTU but the effect direction was still the same.
- S3: The phylogenetic tree was non-informative. We randomly selected 12 OTUs to be aOTUs. We restricted one cluster to have only one aOTU.
- S4: The phylogenetic tree was non-informative. On top of S1, we reversed the effect direction for half of the aOTUs.

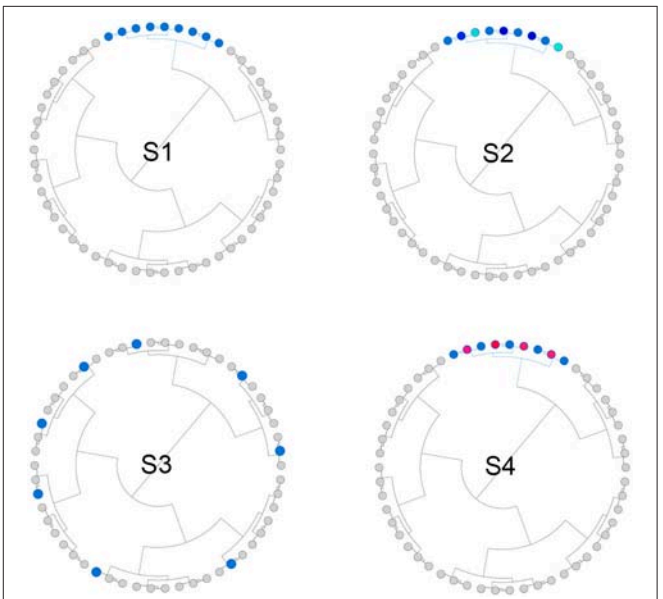


FIGURE 2 | Illustration for the simulation strategy. We simulated both phylogeny-informative scenarios (S1 and S2) and phylogeny-non-informative scenarios (S3 and S4). Blue and red color indicate the direction of the effect while the darkness of the color indicates the magnitude of the effect.

3.1.3. Generating the Outcome Based on the Outcome-Associated OTUs

Denote $\mathcal{A}$ as the set containing the indices of aOTUs, and let x_{ij} be the proportion of OTU j in sample i. We first generated η_i based on the following linear relationship

$$\eta_i = \beta_0 + \sum_{j \in \mathcal{A}} \beta_j x_{ij} \tag{7}$$

For a continuous outcome,

$$y_i = \eta_i + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma_\epsilon^2) \tag{8}$$

For a binary outcome,

$$\begin{aligned} \pi_i &= \frac{e^{\eta_i}}{1 + e^{\eta_i}} \\ y_i &\sim Bernoulli(\pi_i) \end{aligned} \tag{9}$$

We simulated different levels of signal strength (effect size). The signal strength was defined as $\frac{\sqrt{\text{var}(\eta)}}{\sigma_\epsilon}$ for the continuous outcome and $\sum_{j \in \mathcal{A}} \text{var}(\text{x}_j)\beta_j^2$ (x_j denotes the abundance for the jth OTU) for the binary outcome. In the simulation, we investigated a signal strength at 1.0, 1.5, and 2.0 for continuous outcome and 5.0, 10.0, and 20.0 for binary outcome to represent low, medium and high signal strength. The detailed parameter settings for the four scenarios were included in the **Supplementary File**.

3.2. Competing Methods, Model Selection and Evaluation

3.2.1. Competing Methods

We compared the proposed method (SICS) to Lasso, MCP and Elastic Net (Enet), the three sparse regression models without considering the phylogenetic tree. We also compared SICS to a Laplacian-regularized sparse regression model as implemented in *glmgraph* (SLS) (Chen et al., 2015). The Laplacian matrix L was constructed using the same phylogeny-induced correlation matrix C as the adjacency matrix. L was further sparsified to 90% sparsity level to reduce the adverse effects of distantly related OTUs on the outcome prediction. Besides those sparse regression models, we also compared SICS to a representative machine learning method, Random Forest (RF), which has been demonstrated good prediction performance on microbiome data (Pasolli et al., 2016). The parameter settings for the competing methods were shown in **Box 1**.

Box 1 | Parameter settings for competing methods

- Lasso: *glmnet* R package, all parameters were set as the default.
- Elastic Net (Enet): *glmnet* R package. Tuning parameter for l_2 penalty was searched on the grid $\underbrace{\{0, 0.1, 0.2, \cdots, 1\}}_{11}$.
- MCP: *ncvreg* R package, all parameters were set as the default.
- SLS: *glmgraph* R package, the search grid for λ_2 and α were set the same as SICS.
- Random Forest (RF): *randomForest* R package, parameters were set as default.

3.2.2. Model Selection and Evaluation

For SICS, the parameters $(\lambda_1, \lambda_2, \alpha)$ were tuned to achieve optimal model sparsity and phylogenetic depth. Specifically, we searched their best combination over a three-dimensional grids. λ_2 was searched on the grid $\underbrace{\{0, 2^{-5}, 2^{-5+\nu}, 2^{-5+2\nu}, \cdots, 2^5\}}_{12}$, and α on the grid $\underbrace{\{0, 2^{-5}, 2^{-5+\nu}, 2^{-5+2\nu}, \cdots, 2^5\}}_{12}$, $\nu = 1$, while λ_1 was selected from a finer grid on a log scale from the most sparse to a very dense model as implemented in *glmgraph* and *glmnet*.

The best tuning parameter values were selected based on 5-fold cross-validation (CV), where the training samples were randomly divided into 5-folds with 4-folds for model fitting and the remaining fold for testing . We used PMSE (Predicted Mean Square Error) as the CV criterion for a continuous outcome and AUC (Area Under the Curve) for a binary outcome as in Xiao et al. (2018). Once the optimal tuning parameters were selected, we fit the final model using all the training samples and evaluated the prediction on independent testing samples.

To evaluate the prediction performance, we used PMSE ("Brier score" for a binary outcome), which quantifies the *discrepancy* between the predicted and observed values. In addition, we also investigated the R^2, which quantifies the (squared) *correlation* between the predicted and observed values and ranges from 0 (no correlation) to 1 (perfect correlation). Detailed definition of R^2 could be found in Xiao et al. (2018).

Although we focused our evaluation on outcome prediction, variable selection and parameter estimation performance were also investigated to gain more insights about the improved prediction performance of SICS. Variable selection was assessed by sensitivity and specificity, where sensitivity is the true positive rate, i.e., the proportion of aOTUs that are selected, and specificity is true negative rate, i.e., the proportion of irrelevant OTUs that are not selected. The parameter estimation performance was evaluated using MSE (Estimation Mean-Squared Error). Each simulation setting was repeated 50 times and the averages and standard errors of the performance measures were reported.

3.3. Simulation Results

3.3.1. Results for Continuous-Outcome Data

We evaluated the prediction performance in terms of both R^2 and PMSE across different scenarios and signal strengths (**Figure 3**). We observed a general increase in performance for all methods when the signal strength increased. When the phylogenetic tree was informative (Scenario S1 and S2), SICS outperformed other methods substantially with a much larger R^2 and lower PMSE across all levels of signal strength. The improvement of SICS over other methods was more evident when the signal strength decreased, indicating the importance of using the tree prior to pool signals when the signal was weak. Under the weak signal,

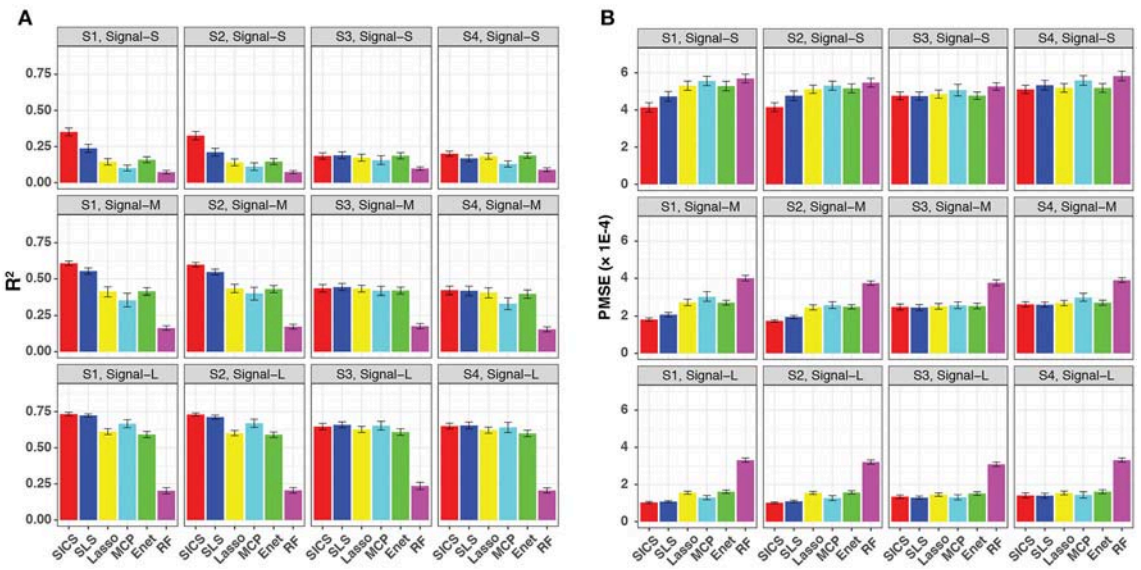


FIGURE 3 | Prediction performance for continuous-outcome simulations across different signal levels and scenarios. Both R^2 **(A)** and PMSE **(B)** were used for evaluation. S1, S2: phylogeny-informative scenarios, and S3, S4: phylogeny-non-informative scenarios; Signal-S, -M, and -L represent weak, medium and strong signals, respectively.

SICS had a clear advantage over SLS, which uses the Laplacian penalty to smooth the coefficients, demonstrating the benefit of using the proposed smoothness penalty that encourages more local smoothing. SICS and SLS were both significantly better than other sparse regression methods and RF across different levels of signal strength. The lower performance of these sparse regression methods was due to their inability to exploit the phylogenetic structure. The improved prediction performance of SICS could be explained by more accurate parameter estimation evidenced by a lower MSE (**Figure S1**) and an increased sensitivity to retain the aOTUs (**Figure S2**). Although the increased sensitivity was at the cost of a slightly lower specificity (**Figure S3**), inclusion of aOTUs was more important than exclusion of non-aOTUs to improve prediction. We also observed that SICS performed similarly in Scenario S1, S2, indicating the robustness of SICS to the variation of the effect size of individual aOTUs as long as the effects are in the same direction.

It should be noted that SICS achieved similar performance as other sparse regression methods in its unfavorable scenarios, when the phylogenetic tree was not informative (Scenarios S3 and S4), demonstrating the robustness of SICS. The comparable performance could be explained by that the additional parameters λ_2, α, which makes MCP and Enet as special cases of SICS.

3.3.2. Results for Binary-Outcome Data

We repeated the same simulations for binary-outcome data and presented the results in **Figure 4**. Compared to the continuous outcome-based simulations, the prediction improvement of SICS was even more striking when the phylogenetic tree was informative (Scenarios S1 and S2). SICS achieved a significantly larger R^2 and smaller Brier Score than other methods across different levels of signal strength. The advantage was even evident when the signal was strong, which was not observed for continuous-outcome data. Overall, a similar trend was observed: SICS had the best performance, followed by SLS under an informative phylogeny; SICS was comparable to other methods for a non-informative phylogeny. The advantage of SICS could be explained by a higher sensitivity of selecting aOTUs (**Figure S4**) at some cost of specificity (**Figure S5**).

3.3.3. Comparison to SLS With Different Sparsity Levels in the Laplacian Matrix

In the above simulation, we adopted a sparsity level of 90% in the Laplacian matrix L for SLS, which generally resulted a satisfactory prediction performance. To further investigate the impact of sparsity level on the prediction performance of SLS, we compared SICS to SLS with different levels of sparsity in L. We tested sparsity levels at 0, 10, 30, 50, 70, and 90% and 0% sparsity indicates no sparsification.

For the continuous-outcome data, SICS consistently outperformed SLS in Scenario S1 & S2 when the signal was weak or medium, and was on par with SLS when the signal was strong (**Figures S6**, **S7**). When the tree was not informative (Scenarios S3, S4), SLS was not sensitive to the sparsity level as expected and the performance was similar to SICS. For binary-outcome data, the performance difference between SICS and SLS was even more striking and SICS performed much better across levels of signal strength when the phylogeny was informative (**Figures S8**, **S9**). We also found that the performance of SLS varied for different levels of sparsity, and SLS generally achieved the best prediction at a sparsity level of 90%. In contrast, SICS did not need to select the optimal sparsity level and had an overall better performance than SLS, regardless of the sparsity level used.

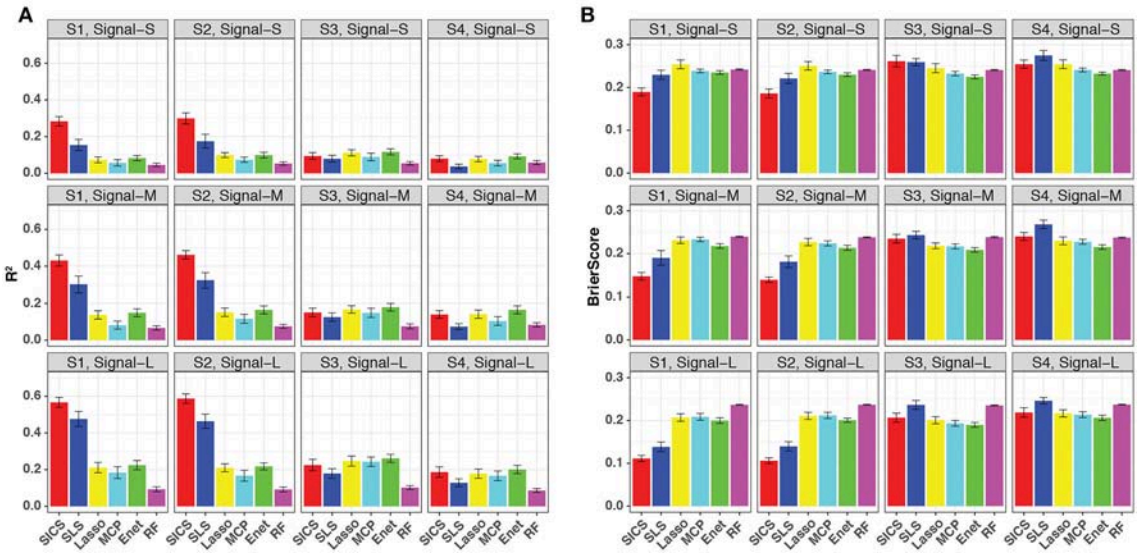


FIGURE 4 | Prediction performance for binary-outcome simulations across different signal levels and scenarios. Both R^2 **(A)** and Brier score **(B)** were used for evaluation. S1, S2: phylogeny-informative scenarios, and S3, S4: phylogeny-non-informative scenarios; Signal-S, -M, and -L represent weak, medium and strong signals, respectively.

4. REAL DATA APPLICATIONS

We applied SCIS to two real microbiome datasets and compared it to the competing methods evaluated in the simulations. We compared to two versions of SLS: SLS without sparsifying L matrix (SLS(0)) and SLS with 90% sparsity level (SLS(0.9)). In addition, we compared to glmmTree, a phylogeny-regularized linear model for dense and clustered microbiome signals (Xiao et al., 2018). The first dataset came from a study of the impact of the long-term dietary pattern on the gut microbiome. We used the caffein intake as the continuous outcome (Wu et al., 2011). The second dataset came from a study of the smoking effect on the human upper respiratory tract microbiome (Charlson et al., 2010). We used the microbiome data from the left side of the throat and treated the smoking status as the binary outcome.

4.1. Caffeine Intake Data

The caffeine intake data was taken from a cross-sectional study of long-term dietary effects on the human gut microbiome in a general population (Wu et al., 2011). The dataset was downloaded from Qiita (https://qiita.ucsd.edu/) with study ID 1011, which consists of 98 samples and 6674 OTUs. We selected the caffeine intake as the outcome of interest since caffeine intake was found to have a significant impact on the gut microbiota (Jaquet et al., 2009). We aimed to predict the caffeine intake based on the OTU abundances. Before applying the prediction methods, we implemented a series of preprocessing steps designed in Xiao et al. (2018) to make the microbiome data more amenable to predictive modeling. First, we removed outlier samples based on an outlier index defined on Bray-Curtis distance and removed rare OTUs with prevalence <10% to reduce the dimensionality of OTUs, leaving 98 samples and 499 OTUs. Second, we normalized OTU raw read counts using GMPR (Chen et al., 2018) followed by a replacement of outlier counts using winsorization at 97% quantile. Third, we transformed the normalized OTU abundance data using square-root transformation to reduce the influence of highly abundant observation. Finally, we applied quantile transformation to the caffeine intake to make it approximately normally distributed.

To have an objective evaluation of the prediction performance, the dataset was randomly divided 50 times into 5 folds each time, among which 4 folds were used for training and the remaining one for testing. In the training set, tuning parameter selection was based on CV as in the simulation. R^2 and PMSE were used as metrics for prediction performance based on the testing set. The results were presented in **Figures 5A,B**. SICS achieved the best performance for caffeine intake prediction as indicated by the highest R^2 and lowest PMSE, followed by Elastic Net, SLS and Random Forest. On the other hand, Elastic Net and SLS, which had data-driven smoothing and prior-driven smoothing, respectively, did improve over Lasso and MCP, which only exploited the model sparsity. To verify whether the improvement of prediction was statistically significant, we performed paired Wilcoxon signed-rank test between SICS and any other methods based on R^2 , PMSE values obtained from the fifty random divisions. SICS achieved significantly higher R^2 , and significantly lower PMSE than any other method ($P<0.05$).

4.2. Smoking Data

The smoking data was from a study of the smoking effect on the human upper respiratory tract microbiome (Charlson et al., 2010). We aimed to predict the smoking status based on the microbiome profile. All the data processing steps were carried out as described in the previous example. After preprocessing,

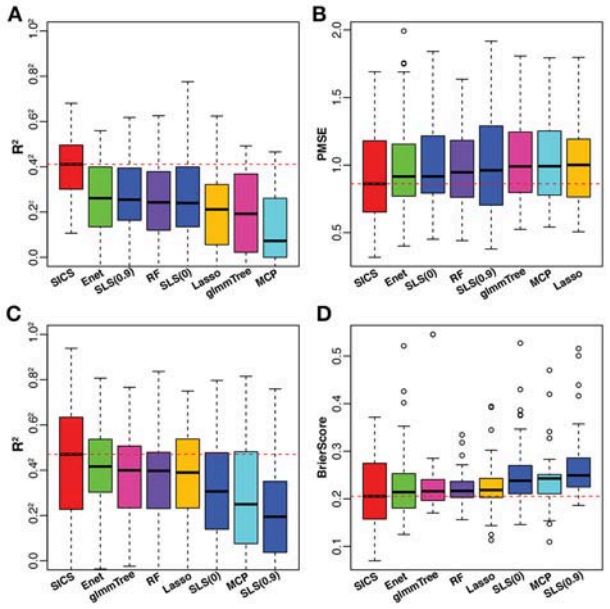


FIGURE 5 | Performance comparison on the caffeine intake data **(A,B)** and smoking data **(C,D)**. The red dashed line indicates the median value of various performance measures for SICS. SLS(0): SLS without sparsification; SLS(0.9): SLS with 90% sparsity level in the Laplacian matrix.

the final dataset consisted of 32 non-smokers and 28 smokers with 174 OTUs. For smoking vs. non-smoking prediction, SICS still achieved the highest R^2 and lowest Brier Score, followed by Elastic Net, glmmTree and Random Forest (**Figures 5C,D**). However, SLS did not improve the prediction performance compared to Lasso and MCP. We also noticed that SLS(0) and SLS(0.9) performed differently (R^2 $P = 0.01$; Brier Score $P = 0.12$). Overall, SICS achieved the best prediction performance for both continuous caffeine intake and dichotomous smoking status.

5. DISCUSSION

The power of a predictive model depends on its capability to exploit the full information in the data, which usually requires domain knowledge. For microbiome data, one unique characteristic is the phylogenetic relationship relating all OTUs, which is important prior information that could be utilized to improve prediction performance. In this paper, we proposed a phylogeny-regularized sparse regression model for capturing sparse and clustered microbiome signals. In the model, a novel phylogeny-based smoothness penalty was designed based on the inverse of phylogeny-induced correlation matrix. We show that such inverse correlation-based smoothness penalty improved over the traditional Laplacian-based smoothness penalty for microbiome applications, due to its local smoothing property as well as the dual smoothing effects (i.e., data-driven and prior-driven smoothing). Moreover, an additional tuning parameter in the smoothness penalty allows our model to capture signals at various phylogenetic depths, further improving its prediction power. We demonstrated the robustness of the proposed method when the tree was not informative or misspecified. A noisy or misspecified tree could be resulted from applying an inappropriate tree construction method or be due to the fact that DNA sequence similarity does not necessarily reflect biological similarity. Interestingly, the proposed method could reduce to Mnet (Huang et al., 2016), which possesses the data-driven smoothing effect.

Similar to other sparse regression models, the proposed method builds on the assumption that the model is sparse: only a few OTUs are associated with the outcome. It is thus expected to be a powerful predictive tool when the signal is sparse. Many diseases have been shown to be associated with a small number of "marker" taxa. For example, in the case of colorectal cancer or arthritis (Scher et al., 2013; Zeller et al., 2014), individual marker taxa were found to be associated to the disease state, whereas effects on the overall composition were very mild. In contrast, other disease states were associated with marked shifts in the overall composition as in the case of obesity and inflammatory bowel disease (Manichanh et al., 2012; Le Chatelier et al., 2013). In such "dense-signal" scenario, sparse regression models including the proposed approach may not work well. Instead, a prediction model based on the global community similarity, such as our recently proposed glmmTree (Xiao et al., 2018), is expected to be more powerful. Exploratory analysis of the microbiome data should be performed before selecting a suitable model.

In the model, we assume a linear relationship between the OTU abundance and the outcome. Although the assumption is usually reasonable after the abundance data is properly normalized and transformed, it may fail to capture complex nonlinear relationship for some applications. Our model can be extended to capture more complex nonlinear effects. The simplest strategy is to apply various transformations, e.g., Box-cox transformation (Sakia, 1992), to the OTU abundance data and selects the best transformation function based on cross-validation. In the case of Box-cox transformation, the power parameter can be treated as another tuning parameter (Xiao and Chen, 2017; Xiao et al., 2018). Alternatively, one could apply an additive model, which is more flexible and allows OTU-specific nonlinear effects (Wood, 2006). However, a larger sample size may be needed to achieve good performance.

Finally, the distribution of OTU abundances is very skewed, and a large number of OTUs are rare and of low-abundance. For these rare OTUs, their sampling variability is very large. Accommodating the sampling error in the predictive model could potentially improve the prediction performance. Jointly modeling the microbiome and the outcome data is thus a promising direction. We leave these extensions as our future work.

AUTHOR CONTRIBUTIONS

JX analyzed the data, wrote the paper, prepared figures and tables, reviewed drafts of the paper. LC analyzed the data, wrote the paper, prepared figures and tables, wrote the software, reviewed drafts of the paper. YY prepared figures and tables, reviewed

drafts of the paper. XZ contributed substantial expertise to improve the paper and revised the paper. JC conceived and designed the experiments, analyzed the data, wrote the paper, wrote the software, prepared figures and tables.

ACKNOWLEDGMENTS

This work was supported by Mayo Clinic Gerstner Family Career Development Awards, Mayo Clinic Center for Individualized Medicine, U01 FD005875, Food and Drug Administration and the National Natural Science Foundation of China (no.61773401 and no.11801571).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.03112/full#supplementary-material

REFERENCES

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

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, e00191–16. doi: 10.1128/mSystems.00191-16

Breheny, P., and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* 5, 232–253. doi: 10.1214/10-AOAS388

Bultman, S. J. (2014). Emerging roles of the microbiome in cancer. *Carcinogenesis* 35, 249–255. doi: 10.1093/carcin/bgt392

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

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. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the illumina hiseq and miseq platforms. *ISME J.* 6: 1621. doi: 10.1038/ismej.2012.8

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., et al. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE* 5:e15216. doi: 10.1371/journal.pone.0015216

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics* 28, 2106–2113. doi: 10.1093/bioinformatics/bts342

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., and Li, H. (2013a). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* 14, 244–258. doi: 10.1093/biostatistics/kxs038

Chen, J., Wright, K., Davis, J. M., Jeraldo, P., Marietta, E. V., Murray, J., et al. (2016). An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med.* 8: 43. doi: 10.1186/s13073-016-0299-7

Chen, L., Liu, H., Kocher, J. P., Li, H., and Chen, J. (2015). glmgraph: an r package for variable selection and predictive modeling of structured genomic data. *Bioinformatics* 31, 3991–3993. doi: 10.1093/bioinformatics/btv497

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). Gmpr: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6:e4600. doi: 10.7717/peerj.4600

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., and Zhao, H. (2013b). A comparison of methods for clustering 16s rrna sequences into otus. *PLoS ONE* 8:e70837. doi: 10.1371/journal.pone.0070837

Chen, X., Johnson, S., Jeraldo, P., Wang, J., Chia, N., Kocher, J. A., et al. (2017). Hybrid-denovo: a *de novo* otu-picking pipeline integrating single-end and paired-end 16s sequence tags. *Gigascience* 7, 1–7. doi: 10.1093/gigascience/gix129

Cho, I., and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat. Rev. Genet* 13, 260–270. doi: 10.1038/nrg3182

Edgar, R. C. (2013). Uparse: highly accurate otu sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Edgar, R. C. (2018). Updating the 97% identity threshold for 16s ribosomal rna otus. *Bioinformatics* 1:5. doi: 10.1093/bioinformatics/bty113

Evans, S. N., and Matsen, F. A. (2012). The phylogenetic kantorovich-rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. B* 74, 569–592. doi: 10.1111/j.1467-9868.2011.01018.x

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Sofw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Garcia, T. P., Muller, S., Carroll, R. J., and Walzem, R. L. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics* 30, 831–837. doi: 10.1093/bioinformatics/btt608

Honda, K., and Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annu. Rev. Immunol.* 30, 759–795. doi: 10.1146/annurev-immunol-020711-074937

Huang, J., Breheny, P., Lee, S., Ma, S., and Zhang, C. (2016). The mnet method for variable selection. *Stat. Sin.* 26, 903–923. doi: 10.5705/ss.202014.0011

Huang, J., Ma, S., Li, H., and Zhang, C. H. (2011). The sparse laplacian shrinkage estimator for high-dimensional regression. *Ann. Stat.* 39, 2021–2046. doi: 10.1214/11-AOS897

Jangi, S., Gandhi, R., Cox, L. M., Li, N., von Glehn, F., Yan, R., et al. (2016). Alterations of the human gut microbiome in multiple sclerosis. *Nat. Commun.* 7:12015. doi: 10.1038/ncomms12015

Jaquet, M., Rochat, I., Moulin, J., Cavin, C., and Bibiloni, R. (2009). Impact of coffee consumption on the gut microbiota: a human volunteer study. *Int. J. Food Microbiol.* 130, 117–121. doi: 10.1016/j.ijfoodmicro.2009.01.011

Kashyap, P. C., Chia, N., Nelson, H., Segal, E., and Elinav, E. (2017). Microbiome at the frontier of personalized medicine. *Mayo Clin. Proc.* 92, 1855–1864. doi: 10.1016/j.mayocp.2017.10.004

Kembel, S. W., Eisen, J. A., Pollard, K. S., and Green, J. L. (2011). The phylogenetic diversity of metagenomes. *PLoS ONE* 6:e23214. doi: 10.1371/journal.pone.0023214

Kinross, J. M., Darzi, A. W., and Nicholson, J. K. (2011). Gut microbiome-host interactions in health and disease. *Genome Med.* 3: 14. doi: 10.1186/gm228

Knights, D., Costello, E. K., and Knight, R. (2011). Supervised classification of human microbiota. *FEMS Microbiol. Rev.* 35, 343–359. doi: 10.1111/j.1574-6976.2010.00251.x

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., et al. (2011). Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet* 13, 47–58. doi: 10.1038/nrg3129

Le Chatelier, E., 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. doi: 10.1038/nature12506

Li, C., and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24, 1175–1182. doi: 10.1093/bioinformatics/btn081

Mallick, H., Ma, S., Franzosa, E. A., Vatanen, T., Morgan, X. C., and Huttenhower, C. (2017). Experimental design and quantitative analysis of microbial community multiomics. *Genome Biol.* 18: 228. doi: 10.1186/s13059-017-1359-z

Manichanh, C., Borruel, N., Casellas, F., and Guarner, F. (2012). The gut microbiota in ibd. *Nat. Rev. Gastroenterol. Hepatol.* 9: 599. doi: 10.1038/nrgastro.2012.152

Martins, E. P., and Hansen, T. F. (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Amer. Nat.* 149, 646–667. doi: 10.1086/286013

Martiny, J. B., Jones, S. E., Lennon, J. T., and Martiny, A. C. (2015). Microbiomes in light of traits: a phylogenetic perspective. *Science* 350:aac9323. doi: 10.1126/science.aac9323

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American gut: an open platform for citizen science microbiome research. *mSystems* 3, e00031–18. doi: 10.1128/mSystems.00031-18

Ning, J., and Beiko, R. G. (2015). Phylogenetic approaches to microbial community classification. *Microbiome* 3: 47. doi: 10.1186/s40168-015-0114-5

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

Peters, B. A., Wu, J., Pei, Z., Yang, L., Purdue, M. P., Freedman, N. D., et al. (2017). Oral microbiome composition reflects prospective risk for esophageal cancers. *Cancer Res.* 77, 6777–6787. doi: 10.1158/0008-5472.CAN-17-1296

Pflughoeft, K. J., and Versalovic, J. (2012). Human microbiome in health and disease. *Annu. Rev. Pathol.* 7, 99–122. doi: 10.1146/annurev-pathol-011811-132421

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2-approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Purdom, E. (2011). Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.* 5, 2326–2358. doi: 10.1214/10-AOAS402

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. doi: 10.1038/nature11450

Randolph, T. W., Zhao, S., Copeland, W., Hullar, M., and Shojaie, A. (2018). Kernel-penalized regression for analysis of microbiome data. *Ann. Appl. Stat.* 12, 540–566. doi: 10.1214/17-AOAS1102

Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., et al. (2014). Subsampled open-reference clustering creates consistent, comprehensive otu definitions and scales to billions of sequences. *PeerJ* 2:e545. doi: 10.7717/peerj.545

Sakia, R. (1992). The box-cox transformation technique: a review. *Statistician* 63, 169–178. doi: 10.2307/2348250

Scher, J. U., Sczesnak, A., Longman, R. S., Segata, N., Ubeda, C., Bielski, C., et al. (2013). Expansion of intestinal prevotella copri correlates with enhanced susceptibility to arthritis. *elife* 2:e01202. doi: 10.7554/eLife.01202

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

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., et al. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 1: 11. doi: 10.1186/2049-2618-1-11

Tanaseichuk, O., Borneman, J., and Jiang, T. (2014). Phylogeny-based classification of microbial communities. *Bioinformatics* 30, 449–56. doi: 10.1093/bioinformatics/btt700

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58, 267–288.

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, 480–484. doi: 10.1038/nature07540

Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F. A., Huttenhower, C., and Jurisica, I. (2011). Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics* 27, 3399–3406. doi: 10.1093/bioinformatics/btr591

Walther-Antonio, M. R., Chen, J., Multinu, F., Hokenstad, A., Distad, T. J., Cheek, E. H., et al. (2016). Potential contribution of the uterine microbiome in the development of endometrial cancer. *Genome Med.* 8: 122. doi: 10.1186/s13073-016-0368-y

Wang, T., and Zhao, H. (2017). Constructing predictive microbial signatures at multiple taxonomic levels. *J. Am. Stat. Assoc.* 112, 1022–1031. doi: 10.1080/01621459.2016.1270213

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. London: Chapman and Hall/CRC.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344

Xiao, J., Cao, H., and Chen, J. (2017). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics* 33, 2873–2881. doi: 10.1093/bioinformatics/btx311

Xiao, J., and Chen, J. (2017). "Phylogeny-based kernels with application to microbiome association studies," in *New Advances in Statistics and Data Science*, eds D. G. Chen, Z. Jin, G. Li, Y. Li, A. Liu, and Y. Zhao (Cham: Springer), 217–237.

Xiao, J., Chen, L., Johnson, S., Yu, Y., Zhang, X., and Chen, J. C. (2018). Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Front. microbiol.* 9:1391. doi: 10.3389/fmicb.2018.01391

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

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 58, 267–288. doi: 10.1214/09-AOS729

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest Statement: 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.

Copyright © 2018 Xiao, Chen, Yu, Zhang and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

TECHNOLOGY REPORT
published: 14 January 2019
doi: 10.3389/fmicb.2018.03336

iVikodak—A Platform and Standard Workflow for Inferring, Analyzing, Comparing, and Visualizing the Functional Potential of Microbial Communities

*Sunil Nagpal, Mohammed Monzoorul Haque, Rashmi Singh and Sharmila S. Mande**

Bio-Sciences R&D Division, TCS Research, Tata Consultancy Services, Pune, India

Background: The objectives of any metagenomic study typically include identification of resident microbes and their relative proportions (taxonomic analysis), profiling functional diversity (functional analysis), and comparing the identified microbes and functions with available metadata (comparative metagenomics). Given the advantage of cost-effectiveness and convenient data-size, amplicon-based sequencing has remained the technology of choice for exploring phylogenetic diversity of an environment. A recent school of thought, employing the existing genome annotation information for inferring functional capacity of an identified microbiome community, has given a promising alternative to Whole Genome Shotgun sequencing for functional analysis. Although a handful of tools are currently available for function inference, their scope, functionality and utility has essentially remained limited. Need for a comprehensive framework that expands upon the existing scope and enables a standardized workflow for function inference, analysis, and visualization, is therefore felt.

Methods: We present iVikodak, a multi-modular web-platform that hosts a logically inter-connected repertoire of functional inference and analysis tools, coupled with a comprehensive visualization interface. iVikodak is equipped with microbial co-inhabitance pattern driven published algorithms along with multiple updated databases of various curated microbe-function maps. It also features an advanced task management and result sharing system through introduction of personalized and portable dashboards.

Results: In addition to inferring functions from 16S rRNA gene data, iVikodak enables (a) an in-depth analysis of specific functions of interest (b) identification of microbes contributing to various functions (c) microbial interaction patterns through function-driven correlation networks, and (d) simultaneous functional comparison between multiple microbial communities. We have bench-marked iVikodak through multiple case studies and comparisons with existing state of art. We also introduce the concept of a public repository which provides a first of its kind community-driven framework for scientific data analytics, collaboration and sharing in this area of microbiome research.

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Yan Zhao,
China University of Mining and Technology, China
Richard Allen White III,
RAW Molecular Systems LLC, United States

****Correspondence:***
Sharmila S. Mande
sharmila.mande@tcs.com

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

Received: *07 October 2018*
Accepted: *24 December 2018*
Published: *14 January 2019*

Citation:
Nagpal S, Haque MM, Singh R and Mande SS (2019) iVikodak—A Platform and Standard Workflow for Inferring, Analyzing, Comparing, and Visualizing the Functional Potential of Microbial Communities. Front. Microbiol. 9:3336. doi: 10.3389/fmicb.2018.03336

Conclusion: Developed using modern design and task management practices, iVikodak provides a multi-modular, yet inter-operable, one-stop framework, that intends to simplify the entire approach toward inferred function analysis. It is anticipated to serve as a significant value addition to the existing space of functional metagenomics.

iVikodak web-server may be freely accessed at https://web.rniapps.net/iVikodak/.

Keywords: inferred functions, 16S metagenome, functional metagenomics, functions of microbial communities, microbiome analysis, visualization, data analyses

INTRODUCTION

Whole Genome Shotgun (WGS) metagenomic DNA sequencing (and subsequent computational analysis of resultant sequence data) helps in not only profiling or cataloging the (microbial) biodiversity characterizing a given habitat, but also enables estimation (of the types and proportions) of various biological functions encoded within the genetic material of microbes resident in that ecological niche (Quince et al., 2017). Multitude of tools and analytical workflows currently exist for WGS driven integrated metagenomics (Keegan et al., 2016; Narayanasamy et al., 2016; White et al., 2017). However, due to high sequencing (and significant downstream computational) costs associated with WGS approach (Bose et al., 2015; Quince et al., 2017; Rossen et al., 2018), initial exploration and estimation of microbial biodiversity (of an environment of interest) is done, in most cases, using amplicon sequencing (Petrosino et al., 2009; Ganju et al., 2016). The latter approach involves PCR amplification and sequencing of a taxonomically informative target genomic marker (e.g., 16S rRNA gene) from the DNA extracted from all microbes present in a given environmental sample. The primary objective of the aforesaid (16S rRNA gene) amplicon-based sequencing has therefore been limited to obtaining quick snapshots of microbial taxonomic diversity in a cost-effective manner. A plethora of bioinformatics tools and standard workflows/ pipelines are currently available for pre-processing and analysing such amplicon sequencing (16S rRNA gene) datasets to meet the said objectives of taxonomic profiling and analyses (Kuczynski et al., 2011; Arndt et al., 2012; McMurdie and Holmes, 2013; Zakrzewski et al., 2017).

A recent school of thought however adds a new dimension to the utility of amplicon sequencing, i.e., "inferring" functional potential of microbial communities "from taxonomic abundance profiles" (Langille et al., 2013). Such inferences are based on the assumption that the pool of genes (and associated functions) in a given microbial community is ultimately a function of the "types and relative abundances" of various microbes (constituting that community). Consequently, given a quantified taxonomic profile corresponding to a given microbial community (residing in a particular environmental niche), it is possible to estimate the functional potential encoded by various microbes constituting the said environment.

A handful of recent methods like PICRUSt (Langille et al., 2013), Tax4Fun (Aßhauer et al., 2015), and Vikodak (Nagpal et al., 2016), have successfully exploited the above mentioned taxa-function inter-relationship for inferring (*in silico*) the functional capacity of microbial communities from their taxonomic profiles (generated through amplicon sequencing). The mentioned methods use distinct algorithms to infer or predict abundances of various functions for a given metagenomic environment in the form of "function abundance matrices." Given that this school of thought is a recent development, the avenue of probing the functional capacity of a metagenomic environment using (16S rRNA gene) amplicon sequencing, has still remained limited to generation of mere "textual matrices" representing the functional abundance for each sample of an environment. A lot of "potential scope" remains unexplored and a standard workflow in the domain of amplicon sequencing driven functional metagenomic analysis is currently lacking. For example, given the availability of various information rich databases like IMG (Markowitz et al., 2012), PATRIC (Wattam et al., 2014), KEGG (Kanehisa and Goto, 2000), that hold prior-collated information about genome-specific functional potentials, it is possible to infer functional correlation based microbe-microbe interaction patterns for a given environment. In addition, functional analysis can also be explored at a granular level to deliver "single pathway or module specific" insights to the researchers. Furthermore, given the availability of established statistical tools and visualization technologies, coupled with multivariate nature of inferred function data, it is possible to define a logical workflow that can not only infer functions, but also perform meta-analyses, statistical comparisons, deep probing and generate meaningful and insightful visualizations.

In the above context, it may be noted that performing such meaningful analysis necessitates end users to garner working knowledge of not only available state-of-art "function prediction" tools (and the input/ output formats, run-time parameters they support), but also an array of statistical and visualization tools, which are additionally required to be implemented for efficient downstream analysis, and hence realize the said workflow for this domain. Needless to say, a platform that enables and (more importantly) automates all of the above mentioned steps is expected to greatly ease the burden on end users. Access to such a framework would enable researchers to efficiently focus on deriving and analysing "functional" insights and

Abbreviations: WGS, Whole Genome Shotgun; PEC, Pathway Exclusion Cut-off; JSD, Jensen-Shannon Divergence; BH, Benjamini-Hochberg; ISFA, Inter Sample Feature Analyzer; PCoA, Principal Coordinate Analysis; ICo, Independent Contributions; CoM, Co-Metabolism; UI, UX, User Interface, User Experience.

subsequently scrutinize the observed trends with respect to related "metadata" and observed taxonomic variation. Such an automated framework would effectively relieve researchers of the mundane nitty-gritty's of (input/ output) data parsing, processing, and visualization support required at almost every stage of analysis, in addition to providing more scope for functional exploration.

Although, the idea of integrating a compendium of tools and utilities to develop such an automated one-stop "infer-compare-visualize" frame-work typically appears more of an e-enablement exercise (driven by IT expertise), it is important to note that the real value add of any such "16S rRNA gene sequencing based" functional annotation framework lies not only in terms of the variety of domain information and functionalities it provides, but also the types of "biologically-relevant" insights it enables for obtaining possible and meaningful answers. The said information, functionalities, and insights may be in terms of (a) *Back-end database*: with respect to the variety, accuracy, and comprehensiveness of its back-end functional-unit (cross)-mappings (b) *Algorithms*: in terms of the types of algorithms and the functional assumptions these algorithms enable for eliciting biologically relevant functional inference(s) (c) *Taxa-function inter-relationships*: in terms of flexibility to back-trace (and visualize in context) the taxonomic sources of the inferred functions and/or deeply probing a pathway or function of interest (d) *Function contribution-based taxonomic relationships/network identification*: derived through co-relation analysis of (inferred) functional capabilities of contributing microbes. Such analysis are expected to enable endusers to additionally narrow down upon, at a microbial (sub) community level, the specific taxonomic drivers behind observed functional patterns or shifts.

Considering the existing state of art in the amplicon sequencing driven functional metagenomic space, we present "iVikodak"—a multi-modular, yet, inter-operable web application framework that provides end users algorithmic options (coupled with updated back-end domain information) for comprehensively inferring the functional potential of microbial communities. At the outset it may be noted that iVikodak represents a significantly upgraded version of Vikodak (Nagpal et al., 2016). iVikodak vastly advances upon the scope and variety of functionalities provided not only in Vikodak, but also other available tools (Langille et al., 2013; Aßhauer et al., 2015; McNally et al., 2018) in the said space. Various functionalities in iVikodak have been intuitively designed and have been e-enabled in formats that simplify for end users the entire approach toward inferred function analysis. Advancements in iVikodak are not limited to the variety of available options for statistical and visual analyses, but also from a user interface (UI) and user-experience (UX) perspective, through development of a well-structured task and data management system. The concept of a public repository (named "ReFDash") which, in addition to hosting pre-generated functional profiles of various environments, provides the research community a frame-work for scientific data collaboration/sharing is also introduced. **Figure 1** provides an overview of iVikodak's functionalities and workflow. Subsequent sections of this paper provide a detailed description of various features of the iVikodak platform. Case-studies highlighting the comparisons with other tools, and utility of specific technical advancements and visualizations incorporated in iVikodak are also provided.

RESULTS AND DISCUSSION

A Comparison of iVikodak With Available Tools and Platforms

As a one-stop "infer-compare-visualize" automated frame-work, iVikodak represents a significant advancement over the textual function abundance matrices generated by the first generation of function inference/prediction tools viz. PICRUSt (Langille et al., 2013), Tax4Fun (Aßhauer et al., 2015), and Vikodak (Nagpal et al., 2016). The prime role of the latter tools is to infer, from end user provided taxonomic profiles, the principal building blocks i.e., information about the types and abundances of various functional units. In an ideal scenario, end users would prefer access to such an inferred or predicted information in terms of varied types of possible functional units (as many as possible). iVikodak, with its updated back-end database (with comprehensive functional unit cross-mapping information) provides, by default, functional annotations in terms of, enzyme copy numbers (EC) (McDonald et al., 2015), TIGRfam (Haft et al., 2003), Pfam (Finn et al., 2014), and COG (Tatusov et al., 2000) categories. Although PICRUSt provides a similar repertoire of information (albeit as textual outputs), iVikodak scores over PICRUSt (and the other two tools) by enabling end users to intuitively compare, probe, and visualize this wealth of information. For instance, the simple automated utility wherein end users are provided queryable tables that not only indicate the subset of bacteria contributing to a chosen function of interest, but also an intuitive (and interactive) visual interface that helps in simultaneously examining the taxonomic and functional context of the same.

In this specific context of comparing iVikodak with other tools of the same genre, it may be noted that a recently published web-utility "Burrito" (McNally et al., 2018) also provides an analogous visualization layout for analysing taxa–function relationships. Using Burrito, users can browse, interactively explore and/or visualize the proportions of individual functions across various samples, along-side the taxa contributing to the said functions. Although Burrito, as a tool, does provide for an automated functional inference cum visualization frame-work besides hosting a decent set of backend parsers and related frontend utilities, it falls short in terms of the following important aspects. Primarily, Burrito's interface is limited to displaying 3 distinct types of visualizations. Prominently, "bar-plot" representations of predicted functions are provided along-side expandable/collapsible "cladograms" that represent corresponding taxonomic hierarchies. These layouts merely enable end users to interactively visualize the types and abundances of various functions (inferred using PICRUSt) in the context of their source taxonomy. In contrast, iVikodak provides end users a comprehensive and inter-operable framework comprised of three logically connected modules, each one of them in turn, having their own repertoire of utilities and multiple

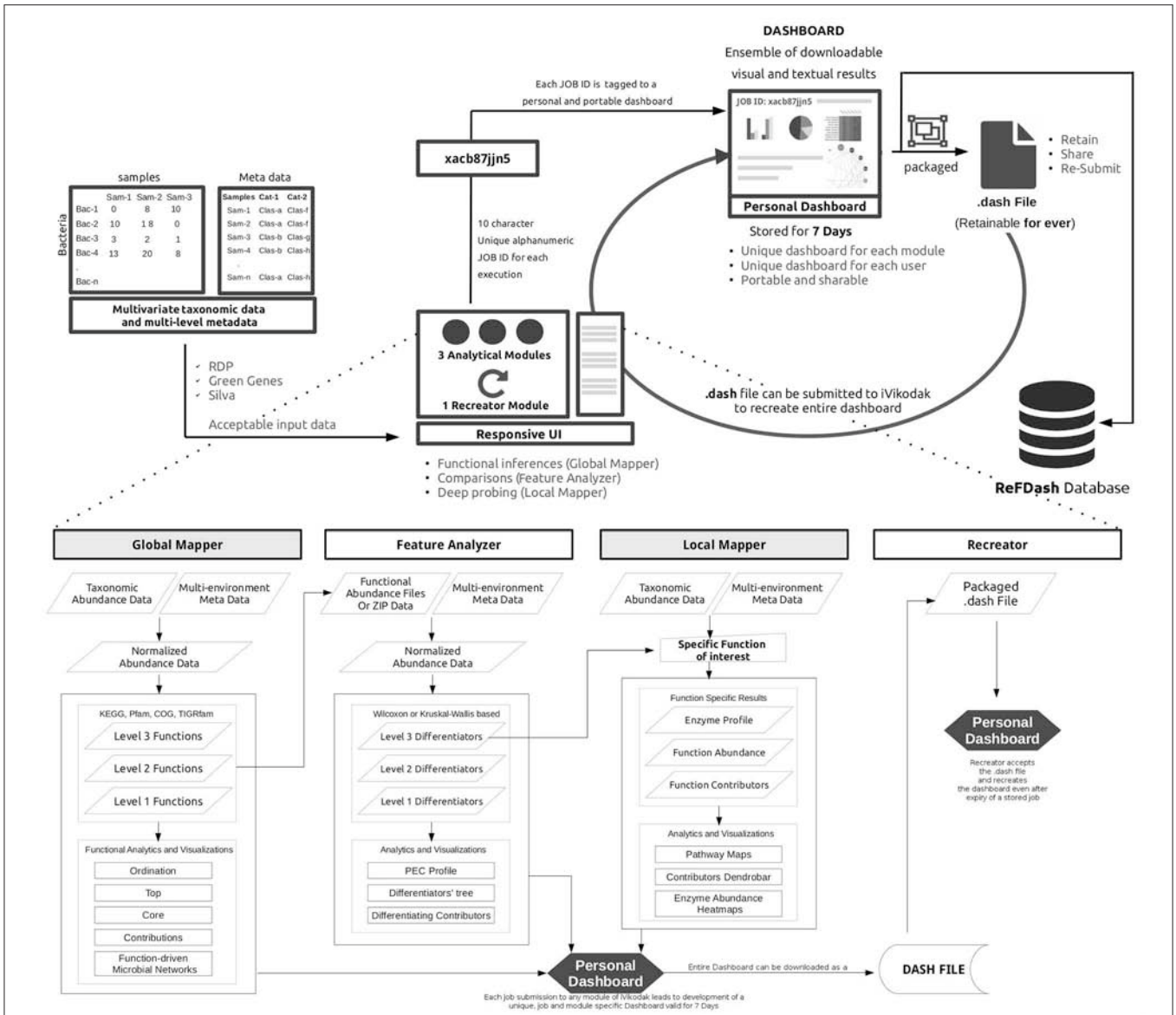


FIGURE 1 | Overview of iVikodak web platform and its structured workflow. An overview of iVikodak's task management and well-structured personalized approach toward 16s rRNA gene sequencing based functional inference, analyses and visualization. Three inter-connected modules of iVikodak ensure comprehensive and meaningful analyses. Each submission to (any module of) iVikodak is tagged to a unique JOB ID, which provides access to a personal and portable dashboard. A dashboard represents an ensemble of "analysable, analyzed, and visualized" results, specific to the chosen module and the type of data uploaded by the end user. The generated dashboards (and associated taxonomic/functional data) can be deposited to the ReFDash repository. This is expected to pave the way for building/populating a community-driven readily accessible database of amplicon sequencing-based functional metagenomics projects and associated data.

types of visualizations. To highlight the array of differences between iVikodak and Burrito, we have used the same dataset (Theriot et al., 2014) that was previously utilized in Burrito for showcasing/ exemplifying its functionalities. This comparison (case study 1) throws light not only on the functionalities available in iVikodak, but also attempts to put forward the vast range of functionalities that were hitherto unavailable (or are not comprehensive enough) in other existing analogous tools including Burrito. **Table 1** provides a comparison of features available in iVikodak, Burrito, Vikodak, PICRUSt and Tax4Fun.

Functionalities Enabling Generation and Visualization of Biologically Meaningful Insights

The visual options provided in iVikodak are not to be construed as a mere e-enablement exercise. For instance, the "*functional networks*" generated (from each of the taxonomic profiles corresponding to one or multiple environments) by the *Global Mapper module* are based on the assumption that microbes co-contributing to specific functions (in a correlated manner) are likely to be interacting (details of various modules provided

TABLE 1 | Comparison of iVikodak with other function inference tools.

Features	iVikodak	Burrito	Vikodak	PICRUSt	Tax4Fun
Multivariate inputs	✓	✓	✓	✓	✓
Default RDP compatibility	✓	×	✓	×	✓
Default Greengenes compatibility	✓	✓	×	✓	×
Default SILVA compatibility	✓	×	✓	×	✓
KEGG, Pfam, COG, TIGRfam inference	✓	✓	×	✓	×
In-depth analysis of pathway of interest	✓	×	✓	×	×
Co-inhabitance based algorithms	✓	×	✓	×	×
Gene Quorum Assumption	✓	×	✓	×	×
Metadata acceptance	✓	✓	✓	×	×
Multiple Categories of Metadata acceptance	✓	×	✓	×	×
Tools for Statistical Comparisons	✓	✓	×	×	×
Graphical Visualizations	✓	✓	×	×	×
- Ordination (PCoA)	✓	×	×	×	×
- Core Functions (Heatmaps)	✓	×	×	×	×
- Top Functions (Grouped Box plots)	✓	×	×	×	×
- Top Functions (Grouped Bar plots)	×	✓	×	×	×
- Differentiating functions (Heatmaps)	✓	×	×	×	×
- Differentiating functions (Cladograms)	✓	×	×	×	×
- Taxa – Function Contribution tree	✓	✓	×	×	×
- Function-driven Networks	✓	×	×	×	×
- Enzyme Abundance Profiles (Heatmaps)	✓	×	×	×	×
3D and Colored KEGG Pathway Maps	✓	×	×	×	×
Task Management (multi-jobs, JOB IDs etc.)	✓	×	×	×	×
Personalized and Portable Dashboards	✓	×	×	×	×

in section Methods, Modules, and Functionalities). Enabling (automated) generation and (interactive) visualization of such networks (and their properties) can potentially help end users in identifying microbial sub-communities that are likely drivers of functional shifts observed between two or more environments. Furthermore, the interactive PCoA ordination plots (in 2D and 3D formats) and the accompanying bar-charts (which visually indicate the proportion of samples in each of the clusters generated during ordination) represent another such functionality (and automated utility) of iVikodak (which tools like Burrito do not provide). It may be noted that during ordination (in iVikodak), samples are clustered based on their (inferred) functional potential, and not as per taxonomic proportions mentioned in the uploaded input profiles.

From an e-enablement perspective, the *range of visualizations* in iVikodak's Global Mapper Module is worth mentioning. "Box-plots" representing abundance of "*top*" (i.e., most abundant) functions, "heat-maps" of functions identified as "*core*" to given environments, 2D/3D ordination plots, function-contribution-based networks are some of the utilities that iVikodak provides. The highlight is that end users can overlay as many types of (available) metadata features over the entire repertoire of visualizations to analyse and download relevant (publication-friendly) images for scientific sharing. Unlike existing tools, right from the step of uploading input data, iVikodak tends to reduce the pre-processing efforts that are typically required to be done by end users. The acceptance of taxonomic profile data generated using any of the three popular taxonomic classification frameworks is one such example.

Apart from identifying and visualizing at various *p*-value thresholds, the set of pathways (or pathway-classes) whose abundances are found to have a statistically significant difference (between two or more environments), the *PEC profile chart* generated by the *inter sample feature analyser (ISFA) module* represents another important functionality that adds value from a biological viewpoint (details in section Methods, Modules, and Functionalities). During the backend process of functional inference, besides providing unfiltered results, iVikodak additionally reports a pathway to be "present" (in a given environment) only when the proportion or the number of its inferred constituent enzymes exceeds at least a minimum quorum of 50%. This threshold is referred to as the "Pathway Exclusion Cut-off" (PEC) threshold (Nagpal et al., 2016). Given that a different functional context may necessitate end users to employ threshold(s) higher than the 50% minimum, iVikodak performs these computations at 5 progressively higher PEC levels, viz. 50, 60, 70, 80, and 90%, and provides a consensus PEC profile chart (in form of a heat-map). This enables users to visually take an informed decision regarding the set of "differentiating" pathways to finally consider (or purge) from their final analysis.

The *local mapper* is a unique module that sets apart iVikodak in comparison to its peers. By enabling end users intending to drill their analysis to the level of individual functional units

that constitute a specific pathway of interest, this module serves as a logical extension to the other two modules of iVikodak (details of the module are provided in section Methods, Modules, and Functionalities). The module provides a contextual platform that facilitates visual analysis of the presence and abundance of various enzymes constituting a given pathway of interest. Typically, end users can employ this module to probe (at a high level of granularity) one or more pathways identified by ISFA module to have a significantly different abundance pattern between the compared environments (e.g., healthy vs. diseased, time-series data, etc.). The facility to generate 3D formatted and colored pathway map(s) is expected to vastly improve the visual (analysis) experience for end users.

Case Study 1: Temporal Observation of Functional Perturbations in Gut Microbiota of Antibiotic Treated Mice

This case-study pertains to available gut microbiome data (from case and time-matched controls) obtained in a longitudinal fashion from mice administered with antibiotics (Theriot et al., 2014). The datasets corresponded to samples obtained at 2 days and at 6 weeks post-treatment with antibiotics (indicated as Abx_Day2, Abx_Day42 for treated samples and Control_Day2, Control_Day42 for control samples).

Figures 2–4 represent a graphical ensemble of (a subset) of key results generated by iVikodak for the aforementioned dataset. The array of results exemplifies the substantially expanded scope/breadth of functionalities available in iVikodak as compared to that in Burrito (McNally et al., 2018) for the same dataset (**Supplementary File 1**). The images in panels 1A and 1B of **Figure 2** depict ordination (JSD-based PCoA) results grouped at two levels viz. as per nature of samples (controls vs. treated) and according to treatment time-points (Day 2 and Day 42 for both controls and treated). It may be noted that the ordination was performed using the functional profiles (of respective samples) that were inferred using Global Mapper module of iVikodak and were automatically (pre)processed for enabling the ordination functionality. While results in panel 1A display the expected trend of segregation between control and treated samples, the clustering profile in panel 1B exhibits a biologically relevant observation, wherein temporal segregation is limited only to the treated samples. Panels 2A and 2B (**Figure 2**) also enable end users to explore, concomitantly visualize, and download a single visual (a readymade box-plot) that captures the pattern of top abundant functions in the provided samples at two different levels of functional hierarchy. Furthermore, iVikodak provides to end users a combined (and easily customizable) heatmap depicting the "core" set of functions (which are "high" as well as "consistently" abundant) in provided sample (and sample classes). Visuals generated for this functionality (depicted in panel 3A and 3B of **Figure 2**) along with that depicted in panel 2 (for the analyzed case-study datasets) enables end users to easily comprehend the relative abundance pattern of various key functions in the analyzed samples.

The (function-driven) taxa correlation networks depicted in panels 4A–E of **Figure 2** unravel quite a few interesting biological insights. These networks clearly depict a state of dysbiosis in gut microbial communities treated with antibiotics. Comparison of networks in panels 4A and 4B visually depicts a marked breakdown of functional interactions between various microbes constituting the gut microbiomes in treated states. It is interesting to note the complete absence of any functional correlations (between any of the members in the bacterial community) 2 days post-antibiotic treatment, and the re-appearance of some interactions 42 days post-treatment. Although this represents an interesting scientific observation (with respect to the immediate impact of antibiotic treatment) whether this represents a true biological event or is it a mere statistical artifact (owing to sample size) remains to be probed further. Overall, it may be noted that these interesting findings (with respect to functional correlation based taxonomic interaction patterns) wouldn't have been obtained using other available analogous tools in this field of research.

The set of images depicted in **Figure 3** enable end users to view (in context) the specific set of functions that display a statistically significant difference in their abundance across the analyzed sample classes. It may be noted that the ISFA module of iVikodak facilitates automated multivariate differentiating feature analyses (both pair-wise as well as multi-class). Of note is ISFA's ability to generate and depict differentiating functions in form of a (three-level) cladogram (representing the functions at all three levels of hierarchy). From the perspective of the present case-study, the cladogram panel in **Figure 3** indicates the relative and significant depletion of various functionalities in antibiotic treated samples. The interactive downloadable taxa-function contribution tables represent a value-add to users intending to further decipher specific functions and/ or taxa of interest.

Taking cues from the results of Global Mapper and ISFA with respect to the depletion of functions related to amino acid metabolism (in antibiotic treated sample class) and also the corresponding observations in both Burrito (McNally et al., 2018) and previous reports (Theriot et al., 2014), we set forth to employ the Local Mapper module of iVikodak to probe this aspect in further detail. For this purpose, we chose to investigate "Arginine biosynthesis" using the same as an example query. Results (depicted in form of intuitive dendrobars in panel 1A–D of **Figure 4**) provide end users a detailed visual insight with respect to the contribution of various microbes (in context of their taxonomic lineage) to this specific function across various classes of the analyzed data. It is apparent from the results (panels 1C and 1D in **Figure 4**) that while contribution of microbes toward this function gets depleted immediately post antibiotic treatment, it gets restored a few weeks post-withdrawal of antibiotic administration. The heat map depicted in panel 2 of **Figure 4** represents the abundance profile of various enzymes constituting this particular function of interest. The heatmap pattern appears to be more or less in sync with previously stated observations. Unlike other existing analogous tools, iVikodak provides readily up-loadable (formatted) KEGG map files that enable end users to generate colored pathway maps visually indicating the difference in enzyme profiles intuitively in a format as depicted in panel 3 of **Figure 4**. The KEGG colored pathway map and heatmap depicting the

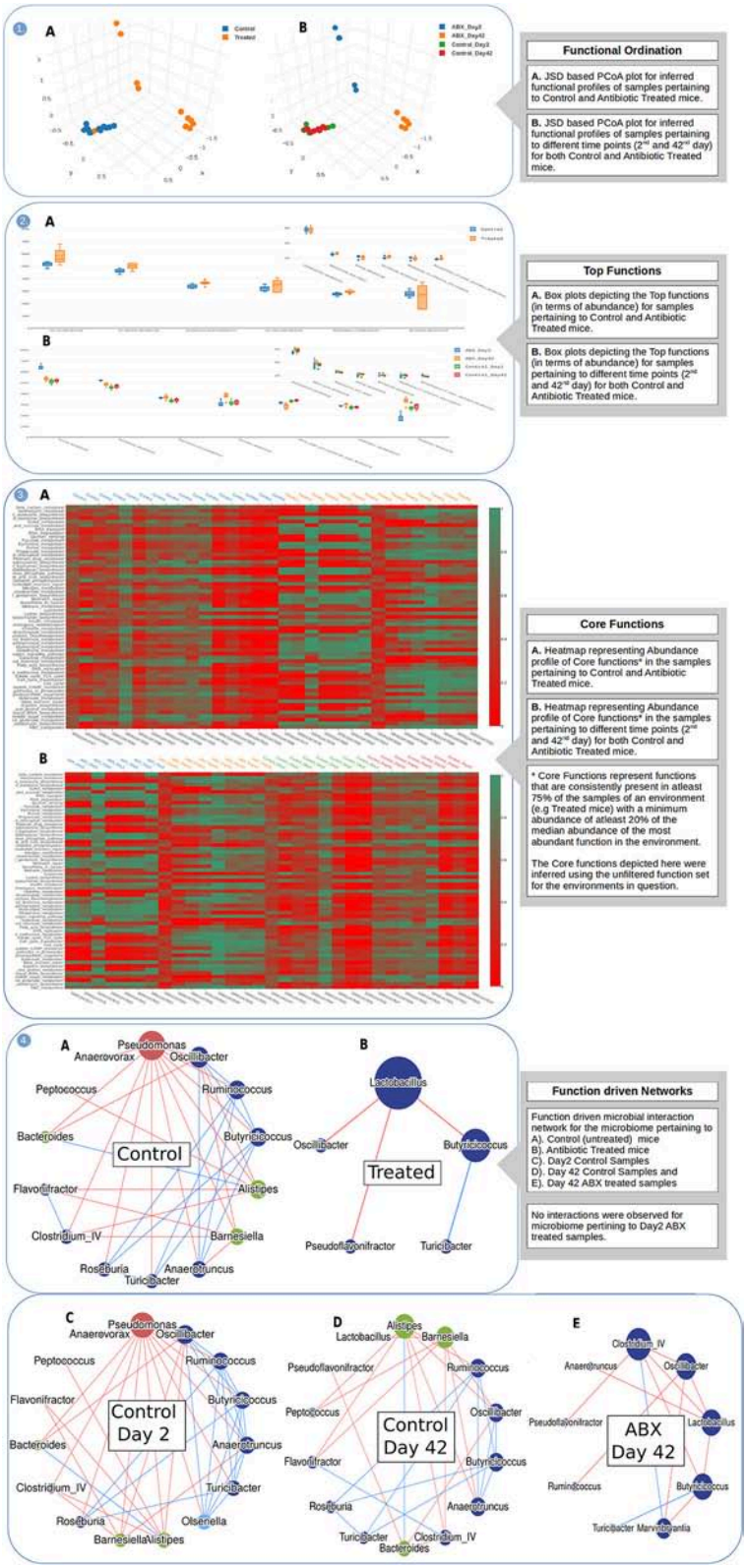


FIGURE 2 | Results of iVikodak's Global Mapper Module for datasets corresponding to case study 1. Plots represent (1) Ordination (2) Top Functions (3) Core Functions (4) Function driven networks aimed at temporal observation of inferred "functional perturbations" in gut microbiota of antibiotic treated mice.

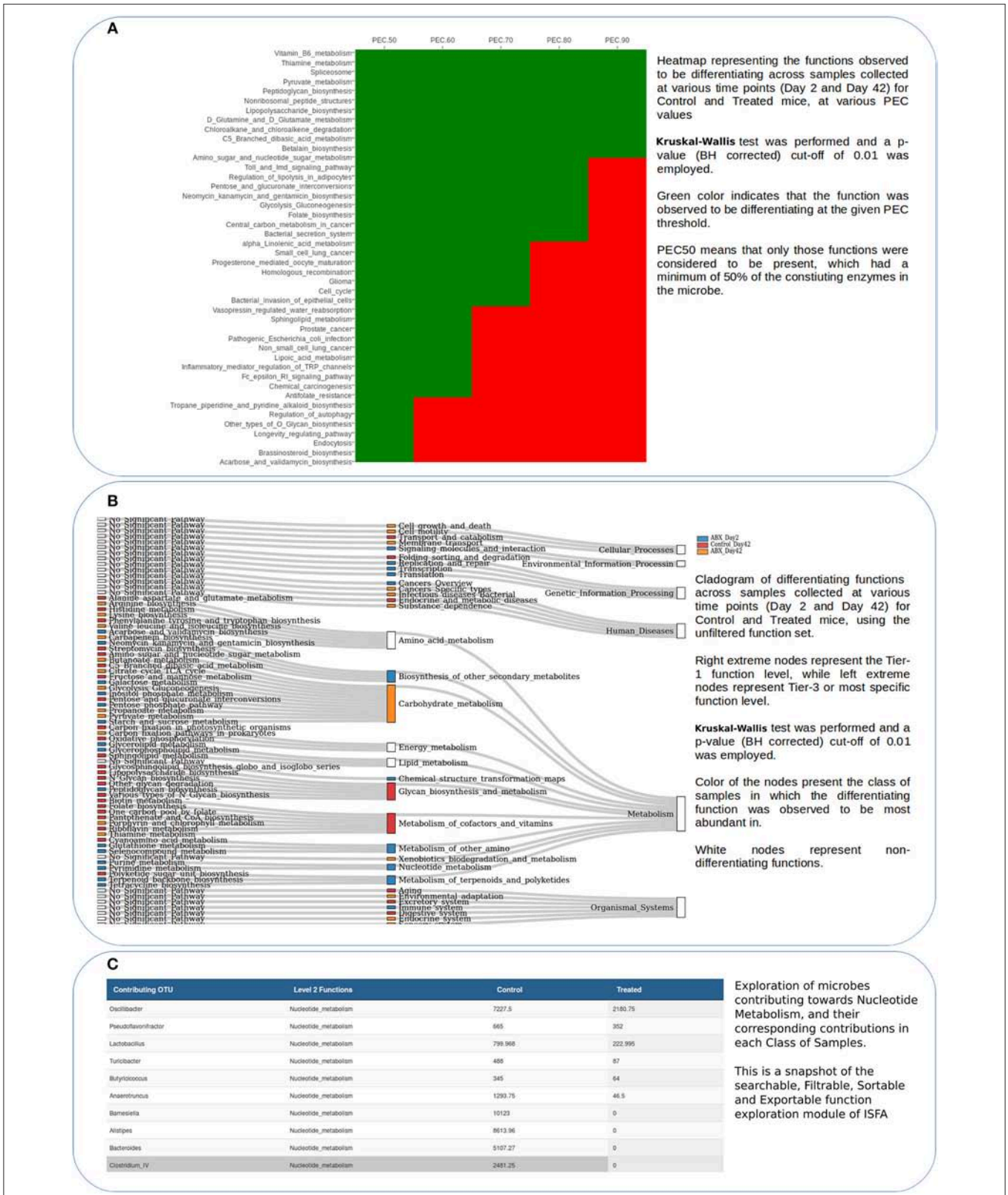
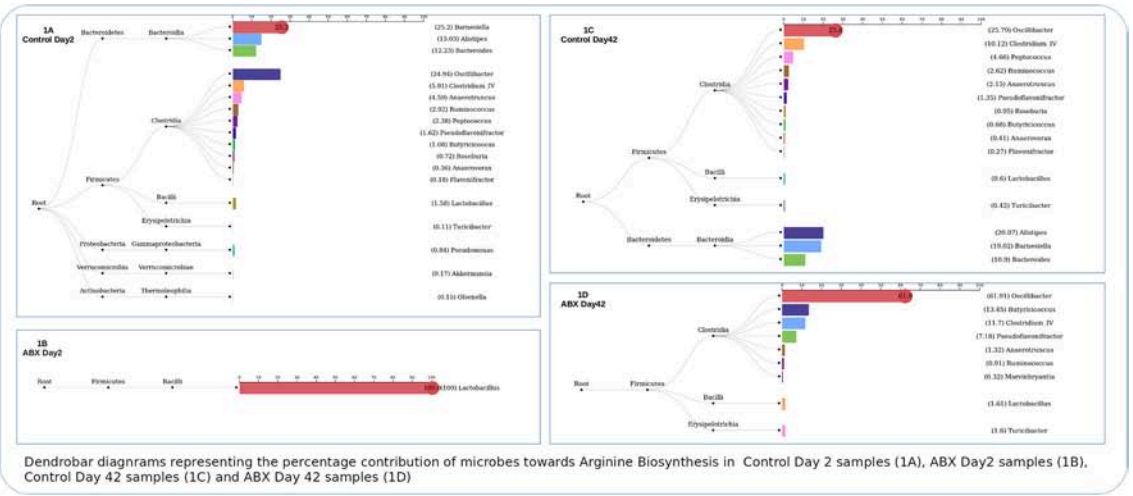
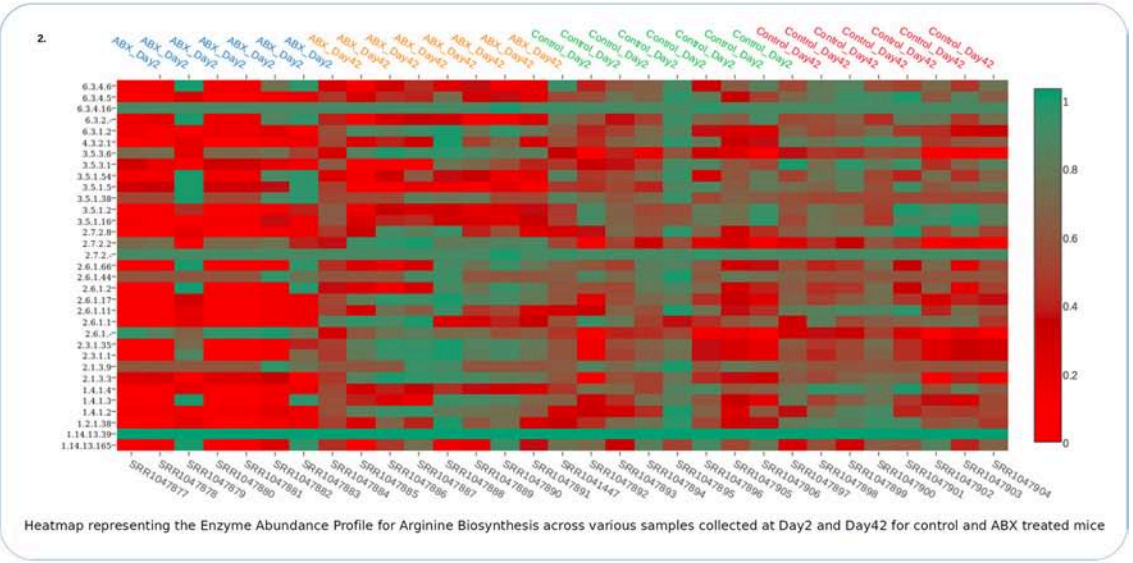


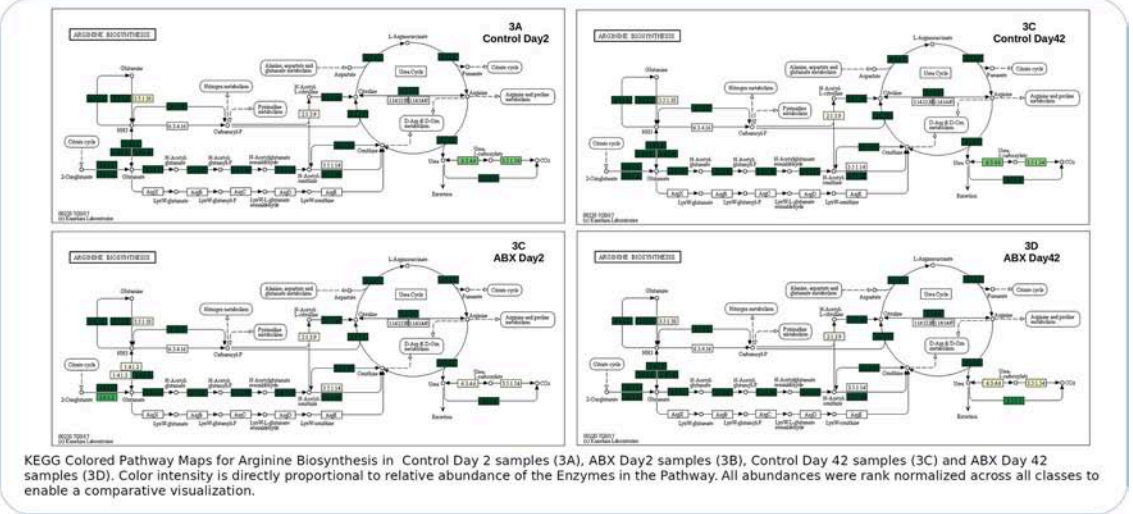
FIGURE 3 | Results of iVikodak's ISFA Module for datasets corresponding to case study 1. Plots represent **(A)** PEC Profile of differentiating functions **(B)** Sankey based cladogram of differentiating functions **(C)** Contributors' profile for differenatiating functions, aimed at temporal observation of inferred "functional perturbations" in gut microbiota of antibiotic treated mice.



Dendrobar diagnrams representing the percentage contribution of microbes towards Arginine Biosynthesis in Control Day 2 samples (1A), ABX Day2 samples (1B), Control Day 42 samples (1C) and ABX Day 42 samples (1D)



Heatmap representing the Enzyme Abundance Profile for Arginine Biosynthesis across various samples collected at Day2 and Day42 for control and ABX treated mice



KEGG Colored Pathway Maps for Arginine Biosynthesis in Control Day 2 samples (3A), ABX Day2 samples (3B), Control Day 42 samples (3C) and ABX Day 42 samples (3D). Color intensity is directly proportional to relative abundance of the Enzymes in the Pathway. All abundances were rank normalized across all classes to enable a comparative visualization.

FIGURE 4 | Results of iVikodak's Local Mapper Module for datasets corresponding to case study 1. Plots for (1) Contributors' dendrobar (2) Pathway specific enzyme profile (3) Colored KEGG Path way Map pertaining to Arginine Biosynthesis for case study 1, qenerated by Local Mapper.

enzyme profile of Arginine Biosynthesis further substantiate the differences between the samples from controls and antibiotic treated mice (as well as the differences at various time points, once again highlighting the extensive metadata handling capacity of iVikodak).

In contrast, the graphic layout depicted in **Supplementary File 1** primarily represents the type of visual analysis that Burrito (as a tool) enables. As seen from this figure, the visualizations generated by Burrito are primarily confined to two broad categories – (a) an interactive grouped/ stacked bar chart that allows end users to visualize the types and proportions of various taxa and functionalities in each uploaded sample (and sample classes), and (b) a collapsible dendrogram that helps in understanding the identified/ predicted functions in their respective hierarchical context. Viewed in the context of the present case study, although the generated charts do provide users an overall sense of important taxonomic and functional differences between the uploaded sample classes, the visualizations remain limited to "only" two types. Hovering over these visualizations pops up information regarding name and proportions of taxa/functions in form of tool-tips. Downloads of visuals highlighting/ showcasing any or all of the displayed information is in form of screen-grabs/SVG's requiring significant post-processing efforts to make them amenable for scientific publication. Overall, even viewed at a mere level of comprehensively meeting e-enablement and automation requirements, a lot of questions, scope and/ or utilities still remain unaddressed by Burrito.

In summary, the results of the above case-study presented in **Figures 2–4** represent the structured and logical connection between the three modules of iVikodak. Commencing from Global Mapper, end users can first probe the data at a community level and then proceed for a comparative analysis of various classes of interest using the next module i.e., ISFA. The deciphered functions of interest from both previous modules lay the perfect context for ultimately performing a detailed visual/ exploratory (and statistical) investigation using the Local Mapper module.

Other Case Studies Highlighting the Functionalities of iVikodak

Five pre-executed jobs are provided as case studies to exemplify various functionalities of all three modules of iVikodak. Job IDs 50a7bef1a5, 998f4e89e5, and d819c619f7 represent "dashboards" for results of Global Mapper, ISFA and Local Mapper for the available Periodontitis microbiome datasets (Aas et al., 2005; Griffen et al., 2012; Souto et al., 2014; Kirst et al., 2015). Briefly, analysis indicates a distinct functional signature common to diseased samples (in contrast to individual-specific pattern in healthy samples). The common set of top core functions identified is in line with functions expected in an oral environment. Distinct changes are observed with respect to the contribution of microbes toward functions that significantly differ between healthy and diseased states. Similarly, Job IDs 5cb3a79c2a and 6c32ef5cda represent results for complex environments (Navarrete et al., 2015; Derycke et al., 2016) and human body sites (Cui et al., 2012; Griffen et al., 2012; Human Microbiome Project Consortium, 2012; Alekseyenko et al., 2013; Botero et al., 2014; Kato et al., 2014; Romero et al., 2014; Xiao et al., 2014) respectively. Details of case studies are provided in **Supplementary File 2**.

IMPLEMENTATION

Overview

IVikodak is a multi-modular framework that enables inference of functional capabilities of a given microbial community, as well as provides an array of automated analytical methods and visualization options for the inferred function data. The three modules in the platform are logically inter-connected to enable automated functional metagenomic analyses in a structured manner, thereby offering a standard working procedure for inferred function driven metagenomics. iVikodak is implemented as a php based web platform developed with modern design practices. The platform employs a job-id and personalized (yet non-intrusive) dashboard-driven task management framework that facilitates seamless end user access to all available utilities and modules without the need for registration and/or the step of sharing personal information (**Figure 1**).

Input Requirements

All modules of iVikodak primarily require two input files (i) Multivariate taxonomic abundance data File (ii) Multi-column metadata of various samples in the taxonomic abundance data. **Supplementary (Video) File 3** and **Supplementary File 4** represent a video tutorial and a schematic representation (respectively) describing the format of a sample taxonomic input data and the corresponding metadata file. Appropriately placed video tutorials, documentation and sample files embedded in tool tips of various modules of the platform, also attempt to provide a succinct guide to the end-users. **Supplementary File 5** contains a listing of various objectives or functions that iVikodak is enabled to provide and perform along with a description of steps (SOP) to be followed for achieving the intended results as well as visualizations.

It may be noted that the existing function inference/prediction tools infer functional abundances using taxonomic input information that have been generated by querying (*in silico*) 16S rRNA gene sequences against "specific reference databases." For example, default implementations of Tax4Fun, Vikodak, and PICRUSt employ (i.e., use as input) taxonomic profiles generated by performing *in silico* comparisons of query sequences against reference taxonomies in SILVA (Pruesse et al., 2007), RDP (Cole et al., 2014), and Greengenes (DeSantis et al., 2006) databases, respectively. Absence of cross-compatibility between tools and the input formats they require therefore presents a challenge to end users. To address this, iVikodak has been enabled to auto-detect, (re-)format, and appropriately process input taxonomic profiles generated using any of the three mentioned reference databases/tools.

Outputs

Each module of iVikodak generates a comprehensive ensemble of "Textual output files" (downloadable as RESULTS.zip from

personal dashboards) and "interactive visualizations/graphs" corresponding to various analytical approaches followed in the module. Said output files/graphs are unique to functionalities of the module used, details pertaining to the same have been described in the following sections of the article.

METHODS, MODULES AND FUNCTIONALITIES

In order to establish a meaningful workflow for performing amplicon sequencing data based functional metagenomic analysis, iVikodak framework incorporates three logically interconnected modules, namely: Global Mapper, Feature Analyzer (ISFA) and Local Mapper. An additional module, named "Recreator" is provided. The latter enables end users to re-create an entire dashboard using the "dashboard specific.dash file" retained by the user post completion of analysis (details in section "Web-front utilities and improvements"). A detailed description of each of three key modules of iVikodak is provided below.

Global Mapper Module

This module enables users to computationally infer (and subsequently visualize and download) the functional potential of one or more microbial environments quantified in terms of the relative abundance of various metabolic pathways. These inferences are obtained by processing (user-supplied) input taxonomic profiles in a series of steps that involve-

(a) Enumerating the repertoire of genes/enzymes/proteins encoded by the set of microbes constituting the provided taxonomic profile
(b) Normalizing this information using 16S and functional gene copy-number information
(c) Cross-mapping this information to other functional units, and
(d) Estimating the relative abundance of various metabolic pathways (in terms of appropriate functional units)

A pre-compiled back-end database is employed by iVikodak's Global Mapper module for performing the first three steps. This database includes copy number and mapping information of over 2,900 enzymes, 15,500 proteins (Pfam: ∼11,200; TIGRfam: ∼4,300), ∼4,600 COGs, and ∼11,000 KOs corresponding to more than 33,000 prokaryotic genomes. This information was collated from IMG (Markowitz et al., 2012) and PATRIC (Wattam et al., 2014) databases. While the first three steps are completed at the back-end (in default automatic mode, as described in Vikodak Nagpal et al., 2016, the final estimation step in Global Mapper module requires end users to provide their preference regarding the functional assumption to proceed with. Choosing "Co-metabolism" (CoM) option results in computing the effective abundance of a metabolic pathway under the assumption that various microbes residing in an environment can pool together the functional units they encode and contribute to the overall functioning of that pathway (in that environment). The other option i.e., "Independent contributions" (ICo), assumes microbes in an environment to be independent functioning entities (with respect to the functional units they encode). Consequently, under this assumption, pathway abundances are independently computed for (and from) each individual microbe (resident in that environment) prior to obtaining their respective sums (which are considered as the effective abundances of respective metabolic pathways).

Besides providing various algorithmic options, the practical utility of the Global Mapper module lies in the variety of (enabled) functional insights that can possibly be queried, obtained, visualized, downloaded, and shared by end users. The option to upload metadata corresponding to each of the samples and directly overlay (and visualize each type of) metadata over the generated results (in an automated manner) is expected to vastly improve the overall visual-experience and aid in showcasing relevant functional insights and differences between environments. Overall, the e-enablement efforts put in behind this module facilitates end users to look beyond simple textual predictions/ inferences, and generate an array of interactive, customizable, publication-friendly graphics (coupled with metadata information of individual environments) in terms of the following functionalities –

(a) Top functions present in these environments (depicted as box plots). These functions may correspond to various pathways or modules or COG/Pfam/TIGRfam classes at various levels of functional hierarchy. Method for "Top Functions" computation is described in **Supplementary File 6**.
(b) Core functions that consistently have a minimum defined abundance in a given environment (depicted as heat-maps). Method for "Core Functions" computation is described in **Supplementary File 6**.
(c) Differentiating functions that exist between two or more environments (based on Wilcoxon-rank sum and Kruskal-Wallis tests) visualized in the context of a cladogram. Method for "Differentiating Functions" computation is described in **Supplementary File 6**.
(d) Ordination Analysis, wherein the inferred functional profiles of the environments under study are subjected to Jensen-Shannon Divergence (JSD) based Principal Coordinate Analysis (Arumugam et al., 2011), and visualized as 2D/3D interactive plots. The sample metadata (e.g., geography, disease status, age, sex, etc.) can also be overlaid on various data points (i.e., microbiome samples) in the generated plot. The latter overlay feature is applicable to all visuals generated by iVikodak.
(e) Function driven Correlation networks, wherein nodes represent microbes and an edge between two microbes indicates that they are (potentially) co-contributing to one or more specific functions. Various centrality measures (degree, betweenness, clustering coefficient, etc.), characterizing the generated networks, are also computed. It may be noted that the relative contribution of individual microbes (constituting a given environment) to an inferred functional unit is obtained by invoking the "independent contributions" algorithm (of Global Mapper) which assumes functional exclusiveness between members of the microbial community. Method used

for generation of "Function-driven Correlation networks" is described in **Supplementary File 6**.

Inter-sample Feature Analyzer (ISFA) Module

This module enables users to perform statistical (multi-class) comparison of functional profiles generated by the Global Mapper module. Wilcoxon rank-sum test and Kruskal-Wallis tests are used for statistical comparison of two or multi-class data, respectively. Both uncorrected as well as Benjamini-Hochberg (BH) corrected (Benjamini and Hochberg, 1995) *p*-values are reported for the features identified as having a significantly different abundance among the compared environments. Users are provided with two modes of operation, namely, rapid and batch mode. The "*rapid*" mode of operation enables (comparative) statistical analysis between functional abundances that have previously been inferred from taxonomic profiles corresponding to two or more environments. These abundances, preferably derived using iVikodak's Global Mapper module (may be generated using any other functional inference tool), are to be provided in typical multivariate data table format to the ISFA module.

The "batch" mode of ISFA, in contrast, works with "Zipped" input data (that is obtained as output from Global Mapper module). The zipped data, containing functional abundance profiles at various Pathway Exclusion Cut-off (PEC) thresholds (Nagpal et al., 2016), enables a consensus-driven differentiating function analysis. PEC threshold based filtering ensures that a pathway is reported as "present" and "functional" only when the (inferred) proportion of its constituent genes/enzymes exceeds a minimum quorum. Interactive cladograms and consensus heat-maps (indicating the list of differentiating functions across all PEC values) are also generated, when the ISFA module is operated in "batch" mode.

The highlight of the ISFA module lies in the *"interactive lists of differentiating functions"* that it automatically generates and displays in a "queryable" format. These lists in form of "filterable, sortable, and exportable" tables represent various inferred functions, bacteria contributing to these functions along with their corresponding quantum of contribution. Given that these lists are "queryable," end users get the flexibility to probe, in real-time, the following two aspects –

(a) Any specific pathway (or pathway-class) of interest, so as to find the list of bacteria contributing to that pathway (along with a comparative view of the proportions in which these bacteria have contributed to various environments or sample classes), and
(b) Any specific bacterial taxon of interest, to estimate its contribution to various pathways (or pathway-classes) in one or more environments.

Local Mapper Module

This module enables a granular-level analysis of user-specified pathway of interest. For a given pathway, end users can probe, visualize, and compare (between environments, and the samples they constitute), the inferred abundances of various enzymes constituting the said pathway. The customizable output, provided in form of a heatmap, in a way depicts the "functional coverage" of any selected pathway across samples or environments. For a given environment (or any other feature provided in metadata), the Local Mapper module also provides an advanced "dendrobar" output format that not only represents the contribution of individual bacteria to the chosen pathway, but also enables users to visualize these bacteria in the context of their taxonomic lineage. As in all other modules, end users are provided drop-down menu options to visualize samples grouped as per metadata. More importantly, iVikodak additionally provides KEGG (Kanehisa and Goto, 2000) color map and KEGG 3D map files corresponding to a user-specified pathway. The latter files are derived from normalized enzyme abundance profiles of the chosen pathway. These ready-to-use pre-formatted output files can be directly uploaded by end users to the KEGG Mapper module of www.genome.jp (KEGG web-server) to generate a colored pathway map and a 3D pathway map (for graphically visualizing relative enzyme abundances).

WEB-FRONT UTILITIES, UI AND UX

In order to provide a seamless user-interface (UI) and user-experience (UX) through a highly interactive and responsive web application, iVikodak uses modern design practices and its front-end employs contemporary state-of-art technologies including bootstrap 3.0, D3.js (Bostock et al., 2011), plotly.js (Plotly Technologies Inc., 2015), cytoscape.js (Franz et al., 2016), in-house java-scripts, etc.

As mentioned earlier, a stand-out feature of iVikodak worth a mention includes incorporation of a personalized and portable "Dashboard" feature. Jobs submitted to iVikodak are tagged to unique "job-ids" that not only help end users to secure or track job status, but also enables them to retrieve, visualize, customize, save, and instantly share the generated results with scientific peers through a personalized "Dashboard." The latter is downloadable as a ".dash" file, which when re-uploaded, seamlessly re-creates the entire dashboard (even post-job ID expiry) for the end user. The dashboard feature in iVikodak is relevant given the present-day trend of open-science and scientific data collaboration or sharing. iVikodak incorporates a "live" task tracking system that provides users a real-time view and access of the following –

- Live status of jobs: Indicating time elapsed, stage of execution i.e., in-progress/completion, etc.,
- Access to intermediate textual results (as an when they are completed) so that users need not wait until the entire dashboard is created.

In addition, post job-completion, the dashboard indicates the total time taken to complete the given job. Apart from enhancing user experience, this feature provides users details pertaining to performance statistics for various submitted jobs. It may be noted that the time taken for various jobs is a function of not only the size of taxonomic abundance data, but also the

scale or the types of metadata provided. It takes approximately 2 min for dashboard generation when a taxonomic abundance data comprises of 500 taxonomic features computed for around 100 microbiome samples with a two-column metadata. It may be noted that the live-tracking system enables users to instantly access (within 10–20 s) the functional profiles that are inferred by iVikodak from the uploaded taxonomic data. It is pertinent to note that the described task management approach and associated feature of dashboard-driven personalization is prominently absent in tools analogous in functionality to iVikodak.

FUTURE DEVELOPMENT AND ENHANCEMENTS

Following are the key enhancements planned for iVikodak –

1. **Algorithmic cross-compatibility**: Currently, functions in iVikodak are inferred using algorithms and methodologies as described in Vikodak (Nagpal et al., 2016). The resultant (inferred) functional information is suitably re-processed (at the backend) and subsequently provided to end users through an interactive visualization interface. As a future enhancement, iVikodak is planned to enable end users to upload functional information inferred using algorithms other than Vikodak e.g., PICRUSt (Langille et al., 2013) and Tax4Fun (Aßhauer et al., 2015). Such uploaded data will also be appropriately re-formatted for enabling respective analyses and visualizations.
2. **Functional Scope:** It is planned to incorporate KEGG modules (Kanehisa and Goto, 2000) and Gene Context based Modules (GCMs) (Bhatt et al., 2018) to the current scope of KO, COG, Pfam and TIGRfam inference.
3. **Custom phylogeny:** The present version of iVikodak accepts taxonomic input data generated using any of the three popular taxonomic classification frame-works (viz. RDP Classifier, Greengenes, and SILVA). Post processing, it currently employs (in its backend) a pre-built taxonomic hierarchy to display respective taxonomic cladograms in various outputs that it generates. The scope of this functionality is planned to be enhanced to enable end users to upload (along with input taxonomic profile data) customized taxonomic as well as functional hierarchies that they wish to use in respective downstream visualizations.
4. **Prediction of Microbe-Disease associations:** In order to derive clinically actionable insights from iVikodak's functional inferences, we intend to enhance functionalities in iVikodak that can appropriately mine disease-specific (inferred) function profiles and decipher in quantifiable terms the association of specific sets of bacteria with a particular disease state (i.e., function driven Human Microbe-Disease Associations, HMDA). A global ensemble of conventional as well as recently developed (validated) algorithms/classifiers like Random Forest (Breiman, 2001), Adaptive Boosting (Freund and Schapire, 1997; Peng et al., 2018), NGRHMDA (Huang et al., 2017), LRLSHMDA (Wang et al., 2017) and other graph theory based approaches (Chen et al., 2017) to extend the utility of microbial co-contribution networks generated by iVikodak, are planned to be included in ISFA module of iVikodak.
5. **ReFDash Database** (Repository of Functional Dashboards): iVikodak's "dashboards" represent a comprehensive ensemble of information, results and visualizations pertaining to the (inferred) functional profiles of one or multiple environments/populations representing one or more microbial communities. Given the fact that each dashboard is unique and can be accessed using personalized JOB IDs, it is possible to create a repository of well annotated dashboards, using public data, as well as through community collaborations.

Given the above context, we intend to create a database (named as ReFDash Repository) with the following objectives –

1. Host pre-generated functional dashboards created using taxonomic profiles (and available metadata) corresponding to various environments.
2. Enable end users to upload, deposit and share their dash-files (or dashboards generated by iVikodak) with other members in the scientific community. The underlying objective is to encourage microbiome researchers to help expand the repertoire of environments (represented by dash files in the ReFDash database). This would ease the process of scientific data collaboration, sharing or even a peer-review process.
3. A further objective is to create a framework that facilitates (automated) "on-demand" (reprocessing and) comparison of functional profiles of a subset of selected environments that are available as pre-generated/ end user deposited dash-files in the ReFDash repository.

Although, the above ideas are under active development, a prototype of the database may be accessed at https://web.rniapps.net/iVikodak/refdash/.

CONCLUSIONS

iVikodak represents an effort to develop a one-stop "infer-analyse-compare-visualize" solution that can assist researchers in deciphering important biological insights with respect to the functional potential of microbial communities based on 16S rRNA gene sequencing datasets. The modular (yet inter-operable) framework of iVikodak intends to lay-down a standard workflow for inferred functional metagenomics. It can facilitate end users to concomitantly infer, statistically analyze, and compare multiple microbial communities, and in the process generate a plethora of intuitive self-explanatory visual outputs in an automated fashion. The ISFA and Local Mapper modules of iVikodak are logical extensions of the Global Mapper module and their expanded scope now enable end users to automate the statistical comparison between the functional potential of multiple environments (and their corresponding classes or metadata). The planned development (and eventual) linkage to ReFDash Repository represents the broad vision behind this work, and it is anticipated that iVikodak will add a

significant value to the existing space of inferred function driven metagenomics space.

AVAILABILITY AND REQUIREMENTS

1. **Project name**: iVikodak
2. **Home page**: https://web.rniapps.net/iVikodak/
3. **Tutorials:** Each module has dedicated video tutorials. Comprehensive documented tutorials may be accessed at: https://web.rniapps.net/iVikodak/tutorials.documentation/
4. **Operating system(s):** Web platform compatible with all operating systems
5. **Browser requirement(s):** WebGl enabled for viewing 3D plots

For inquiries and general discussions, please contact sunil.nagpal@tcs.com, mm.haque@tcs.com or sharmila.mande@tcs.com.

AVAILABILITY OF DATA AND MATERIALS

Publicly available datasets were used in all the case studies described in this manuscript. A summary of their source(s) has been provided in **Supplementary File 7**. In addition, the taxonomic abundance (data) matrices, along-with their meta-data, for all the datasets used in this manuscript (in addition to other datasets) have been populated in the ReFDash database (https://web.rniapps.net/iVikodak/refdash/). **Supplementary File 8** is an archive comprising of the mentioned datasets. This archive also includes (a) various scripts that enable regeneration of R plots from the output matrices generated by iVikodak, as well as (b) the code base for analytical work-flows for identification and generation of visualizations corresponding to top and core functions.

AUTHOR CONTRIBUTIONS

SN conceived, designed and developed the method, algorithm and web platform for iVikodak. SN and RS mined back end data and performed case studies. SN, MH, and SM conceived idea for ReFDash. SN, RS, and MH mined data for ReFDash. MH, SN, RS, and SM prepared the manuscript. All authors have reviewed and validated the platform and manuscript.

FUNDING

The authors declare that this study received funding in form of monthly remuneration (salary) for the authors, from TCS Ltd. The funder had the role of a promoter of fundamental research in BioSciences Research Division of TCS Research. This study was an outcome of one of such fundamental research efforts. If anyone intends to use the outcomes of this research for commercial goals, TCS Ltd shall hold the rights for such commercial relationships with respect to iVikodak.

ACKNOWLEDGMENTS

Authors would like to thank colleagues in BioSciences Research group (of TCS Research) for their help in testing the iVikodak platform.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.03336/full#supplementary-material

REFERENCES

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., and Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* 43, 5721–5732. doi: 10.1128/JCM.43.11.5721-5732.2005

Alekseyenko, A. V., Perez-Perez, G. I., De Souza, A., Strober, B., Gao, Z., Bihan, M., et al. (2013). Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome* 1:31. doi: 10.1186/2049-2618-1-31

Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A. C., Cruz, J. A., et al. (2012). METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Res.* 40, W88–W95. doi: 10.1093/nar/gks497

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, 174–180. doi: 10.1038/nature09944

Aßhauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 31, 2882–2884. doi: 10.1093/bioinformatics/btv287

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

Bhatt, V., Mohapatra, A., Anand, S., Kuntal, B. K., and Mande, S. S. (2018). FLIM-MAP: gene context based identification of functional modules in bacterial metabolic pathways. *Front. Microbiol.* 9:2183. doi: 10.3389/fmicb.2018.02183

Bose, T., Haque, M. M., Reddy, C., and Mande, S. S. (2015). COGNIZER: a framework for functional annotation of metagenomic datasets. *PLoS ONE* 10:e0142102. doi: 10.1371/journal.pone.0142102

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 data-driven documents. *IEEE Trans. Visual. Comput. Graph.* 17, 2301–2309. doi: 10.1109/TVCG.2011.185

Botero, L. E., Delgado-Serrano, L., Cepeda, M. L., Bustos, J. R., Anzola, J. M., Del Portillo, P., et al. (2014). Respiratory tract clinical sample selection for microbiota analysis in patients with pulmonary tuberculosis. *Microbiome* 2:29. doi: 10.1186/2049-2618-2-29

Breiman, L. (2001). Random forests. *Mach Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

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

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

Cui, Z., Zhou, Y., Li, H., Zhang, Y., Zhang, S., Tang, S., et al. (2012). Complex sputum microbial composition in patients with pulmonary tuberculosis. *BMC Microbiol.* 12:276. doi: 10.1186/1471-2180-12-276

Derycke, S., Meester, N. D., Rigaux, A., Creer, S., Bik, H., Thomas, W. K., et al. (2016). Coexisting cryptic species of the Litoditis marina complex (Nematoda) show differential resource use and have distinct microbiomes with high intraspecific variability. *Mol. Ecol.* 25, 2093–2110. doi: 10.1111/mec.13597

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

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32, 309–311. doi: 10.1093/bioinformatics/btv557

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Syst. Sci.* 55, 119–139. doi: 10.1006/jcss.1997.1504

Ganju, P., Nagpal, S., Mohammed, M. H., Kumar, P. N., Pandey, R., Natarajan, V. T., et al. (2016). Microbial community profiling shows dysbiosis in the lesional skin of Vitiligo subjects. *Sci. Rep.* 6:18761. doi: 10.1038/srep18761

Griffen, A. L., Beall, C. J., Campbell, J. H., Firestone, N. D., Kumar, P. S., Yang, Z. K., et al. (2012). Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* 6, 1176–1185. doi: 10.1038/ismej.2011.191

Haft, D. H., Selengut, J. D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res.* 31, 371–373. doi: 10.1093/nar/gkg128

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

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

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kato, T., Fukuda, S., Fujiwara, A., Suda, W., Hattori, M., Kikuchi, J., et al. (2014). Multiple omics uncovers host-gut microbial mutualism during prebiotic fructooligosaccharide supplementation. *DNA Res.* 21, 469–480. doi: 10.1093/dnares/dsu013

Keegan, K. P., Glass, E. M., and Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol Biol.* 1399, 207–233. doi: 10.1007/978-1-4939-3369-3_13

Kirst, M. E., Li, E. C., Alfant, B., Chi, Y.-Y., Walker, C., Magnusson, I., et al. (2015). Dysbiosis and alterations in predicted functions of the subgingival microbiome in chronic periodontitis. *Appl. Environ. Microbiol.* 81, 783–793. doi: 10.1128/AEM.02712-14

Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., and Knight, R. (2011). Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr. Protoc. Bioinform.* Chapter, Unit10.7. doi: 10.1002/0471250953.bi1007s36

Langille, M. G. I., 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

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122. doi: 10.1093/nar/gkr1044

McDonald, A. G., Boyce, S., and Tipton, K. F. (2015). "Enzyme classification and nomenclature," in *eLS* (Chichester: John Wiley & Sons, Ltd). doi: 10.1002/9780470015902.a0000710.pub3

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8:e61217. doi: 10.1371/journal.pone.0061217

McNally, C. P., Eng, A., Noecker, C., Gagne-Maynard, W. C., and Borenstein, E. (2018). BURRITO: an interactive multi-omic tool for visualizing taxa-function relationships in microbiome data. *Front. Microbiol.* 9:365. doi: 10.3389/fmicb.2018.00365

Nagpal, S., Haque, M. M., and Mande, S. S. (2016). Vikodak - a modular framework for inferring functional potential of microbial communities from 16S metagenomic datasets. *PLoS ONE* 11:e0148347. doi: 10.1371/journal.pone.0148347

Narayanasamy, S., Jarosz, Y., Muller, E. E. L., A., Heintz-Buschart, Herold, M., Kaysen, A., et al. (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* 17:260. doi: 10.1186/s13059-016-1116-8

Navarrete, A. A., Tsai, S. M., Mendes, L. W., Faust, K., de Hollander, M., Cassman, N. A., et al. (2015). Soil microbiome responses to the short-term effects of Amazonian deforestation. *Mol. Ecol.* 24, 2433–2448. doi: 10.1111/mec.13172

Peng, L.-H., Yin, J., Zhou, L., Liu, M.-X., and Zhao, Y. (2018). Human microbe-disease association prediction based on adaptive boosting. *Front. Microbiol.* 9:2440. doi: 10.3389/fmicb.2018.02440

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

Plotly Technologies Inc. (2015). *Collaborative Data Science*. Plotly Technologies Inc. Available online at: https://plot.ly

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196. doi: 10.1093/nar/gkm864

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 833–844. doi: 10.1038/nbt.3935

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Bieda, J., et al. (2014). The vaginal microbiota of pregnant women who subsequently have spontaneous preterm labor and delivery and those with a normal delivery at term. *Microbiome* 2:18. doi: 10.1186/2049-2618-2-18

Rossen, J. W. A., Friedrich, A. W., and Moran-Gilad, J. (2018). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin. Microbiol. Infect.* 24, 355–360. doi: 10.1016/j.cmi.2017.11.001

Souto, R., Silva-Boghossian, C. M., and Colombo, A. P. V. (2014). Prevalence of *Pseudomonas aeruginosa* and Acinetobacter spp. in subgingival biofilm and saliva of subjects with chronic periodontal infection. *Braz. J. Microbiol.* 45, 495–501. doi: 10.1590/S1517-83822014000200017

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

Theriot, C. M., Koenigsknecht, M. J., Carlson, P. E., Hatton, G. E., Nelson, A. M., Li, B., et al. (2014). Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nat. Commun.* 5:3114. doi: 10.1038/ncomms4114

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42, D581–D591. doi: 10.1093/nar/gkt1099

White, R. A., Brown, J., Colby, S., Overall, C. C., Lee, J.-Y., Zukcer, J., et al. (2017). ATLAS (Automatic Tool for Local Assembly Structures) - a comprehensive infrastructure for assembly, annotation, and genomic binning of metagenomic and metatranscriptomic data. *PeerJ* 5:e2843ve21. doi: 10.7287/peerj.preprints.2843v1

Xiao, S., Fei, N., Pang, X., Shen, J., Wang, L., Zhang, B., et al. (2014). A gut microbiota-targeted dietary intervention for amelioration of chronic inflammation underlying metabolic syndrome. *FEMS Microbiol. Ecol.* 87, 357–367. doi: 10.1111/1574-6941.12228

Zakrzewski, M., Proietti, C., Ellis, J. J., Hasan, S., Brion, M.-J., Berger, B., et al. (2017). Calypso: a user-friendly web-server for mining and visualizing microbiome–environment interactions. *Bioinformatics* 33, 782–783. doi: 10.1093/bioinformatics/btw725

Conflict of Interest Statement: 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.

Copyright © 2019 Nagpal, Haque, Singh and Mande. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

TECHNOLOGY REPORT
published: 24 January 2019
doi: 10.3389/fmicb.2018.03349

SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline

Javier Tamames and Fernando Puente-Sánchez*

Department of Systems Biology, Spanish Center for Biotechnology, CSIC, Madrid, Spain

The improvement of sequencing technologies has facilitated generalization of metagenomic sequencing, which has become a standard procedure for analyzing the structure and functionality of microbiomes. Bioinformatic analysis of sequencing results poses a challenge because it involves many different complex steps. SqueezeMeta is a fully automatic pipeline for metagenomics/metatranscriptomics, covering all steps of the analysis. SqueezeMeta includes multi-metagenome support that enables co-assembly of related metagenomes and retrieval of individual genomes via binning procedures. SqueezeMeta features several unique characteristics: co-assembly procedure or co-assembly of unlimited number of metagenomes via merging of individual assembled metagenomes, both with read mapping for estimation of the abundances of genes in each metagenome. It also includes binning and bin checking for retrieving individual genomes. Internal checks for the assembly and binning steps provide information about the consistency of contigs and bins. Moreover, results are stored in a MySQL database, where they can be easily exported and shared, and can be inspected anywhere using a flexible web interface that allows simple creation of complex queries. We illustrate the potential of SqueezeMeta by analyzing 32 gut metagenomes in a fully automatic way, enabling retrieval of several million genes and several hundreds of genomic bins. One of the motivations in the development of SqueezeMeta was producing a software capable of running in small desktop computers and thus amenable to all users and settings. We were also able to co-assemble two of these metagenomes and complete the full analysis in less than one day using a simple laptop computer. This reveals the capacity of SqueezeMeta to run without high-performance computing infrastructure and in absence of any network connectivity. It is therefore adequate for *in situ*, real time analysis of metagenomes produced by nanopore sequencing. SqueezeMeta can be downloaded from https://github.com/jtamames/SqueezeMeta.

Keywords: binning, metagenomics, MinION, RNAseq, software

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Alfredo Ferro,
Università degli Studi di Catania, Italy
Yu-Wei Wu,
Taipei Medical University, Taiwan

****Correspondence:***
Javier Tamames
jtamames@cnb.csic.es

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

Received: *19 September 2018*
Accepted: *31 December 2018*
Published: *24 January 2019*

Citation:
Tamames J and Puente-Sánchez F (2019) SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline. Front. Microbiol. 9:3349. doi: 10.3389/fmicb.2018.03349

INTRODUCTION

The improvement of sequencing technologies has permitted the generalization of metagenomic sequencing, which has become standard procedure for analyzing the structure and functionality of microbiomes. Many novel bioinformatic tools and approaches have been developed to deal with the vast numbers of short read sequences produced by a metagenomic experiment. Aside from the simply overwhelming amount of data, a metagenomic analysis is a complex task comprising several non-standardized steps, involving different software tools whose results are often not directly compatible.

Lately, the development of highly portable sequencers, especially those based on nanopore technologies (Deamer et al., 2016), has facilitated *in situ* sequencing in scenarios where the need to obtain quick results is paramount, for instance clinical scenarios of disease control or epidemics (Quick et al., 2015, 2016). Metagenomic sequencing has also been performed *in situ*, for instance in oceanographic expeditions in the Antarctic ice (Lim et al., 2014; Johnson et al., 2017), illustrating the growing capability of producing sequences right away in sampling campaigns. This will enable informed planning of upcoming sampling experiments according to the results found in previous days. We foresee that this kind of application will be increasingly used in the near future. Therefore, bioinformatic analysis should be performed in a very short time span (hours), and be amenable to lightweight computing infrastructure.

A standard metagenomic pipeline involves read curation, assembly, gene prediction, and functional and taxonomic annotation of the resulting genes. Several pipelines have been created to automate most of these analyses (Li, 2009; Arumugam et al., 2010; Glass and Meyer, 2011; Abubucker et al., 2012; Eren et al., 2015; Kim et al., 2016). However, they differ in terms of capacities and approaches. One of the most important differences is whether or not the assembly step is needed. Some platforms skip assembly and, consequently, gene prediction and rely instead on direct annotation of the raw reads. Nevertheless, there are several drawbacks of working with raw reads: since this is based on homology searches for millions of sequences against huge reference databases, it usually requires very large CPU usage. Especially for taxonomic assignment, the reference database must be as complete as possible to minimize errors (Pignatelli et al., 2008). Furthermore, sequences are often too short to produce accurate assignments (Wommack et al., 2008; Carr and Borenstein, 2014).

Assembly, however, is advisable because it can recover larger fragments of genomes, often comprising many genes. Having the complete sequence of a gene and its context makes its functional and taxonomic assignment much easier and more reliable. The drawback of assembly is the formation of chimeras because of misassembling parts of different genomes, and the inability to assemble some of the reads, especially the ones from low-abundance species. The fraction of non-assembled reads depends on several factors, especially sequencing depth and microbiome diversity, but it is usually low (often below 20%). Recently, some tools have been developed to reassemble the portion of reads not assembled in the first instance, increasing the performance of this step (Hitch and Creevey, 2018). Co-assembling related metagenomes can also alleviate this problem significantly, as we will illustrate in the results section.

Assembly is also advisable because it facilitates the recovery of quasi-complete genomes via binning methods. The retrieval of genomes is a major step forward in the study of a microbiome, since it enables linking organisms and functions, thereby contributing to a much more accurate ecologic description of the community's functioning. It is possible, for instance, to determine the key members of the microbiome (involved in particularly important functions), to infer potential interactions between members (for instance, looking for metabolic complementation), and to advance in the understanding of the effect of ecologic perturbations.

The best strategy for binning is co-assembly of related metagenomes. By comparing the abundance and composition of the contigs in different samples, it is possible to determine which contigs belong to the same organism: these contigs have similar oligonucleotide composition, similar abundances in individual samples, and a co-varying pattern between different samples. In this way, it is possible to retrieve tens or hundreds of genomic bins with different levels of completion that can be used as the starting point for a more in-depth analysis of the microbiome's functioning.

SqueezeMeta is a fully automatic pipeline for metagenomics/metatranscriptomics, covering all steps of the analysis. It includes multi-metagenome support allowing co-assembly of related metagenomes and the retrieval of individual genomes via binning procedures.

A comparison of the capabilities of SqueezeMeta and other pipelines is shown in **Table 1**. Most current pipelines do not include support for co-assembling and binning, while some permit importing external binning results to display the associated information.

SqueezeMeta offers several advanced characteristics that make it different to existing pipelines, for instance:

1. Co-assembly procedure coupled with read mapping for the estimation of the abundances of individual genes in each metagenome.
2. An alternative co-assembly approach enabling the processing of an unlimited number of metagenomes via merging of individual metagenomes.
3. Support for nanopore long reads.
4. Binning and bin checking for retrieving individual genomes.
5. Internal checks for the taxonomic annotation of contigs and bins.
6. Metatranscriptomic support via mapping of cDNA reads against reference metagenomes, or via co-assembly of metagenomes and metatranscriptomes.
7. Inclusion of MySQL database for storing results, where they can be easily exported and shared and inspected anywhere using a web interface.

We have designed SqueezeMeta to be able to run in scarce computer resources, as expected for *in situ* metagenomic

TABLE 1 | Features of different metagenomic analysis pipelines, in comparison to SqueezeMeta.

	MG-Rast (Meyer et al., 2008)	Anvio (Eren et al., 2015)	Smash community (Arumugam et al., 2010)	Humann (Abubucker et al., 2012)	fmap (Kim et al., 2016)	MetaWrap (Uritskiy et al., 2018)	Samsa2 (Westreich et al., 2018)	IMP (Narayanasamy et al., 2016)	Squeeze Meta
Assembly	No	No	Yes	No	No	Yes	No	Yes	Yes
Data source	Reads or contigs	Contigs	Contigs	Reads	Reads or contigs	Contigs	Reads (RNA)	Reads	Reads
Gene prediction	Yes	Yes	Yes	No	No	No	No	Yes	Yes
Function assignment	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
RNA assignment	Yes	Yes	No	No	No	No	Yes	Yes	Yes
Taxonomic assignment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gene abundances	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
Metagomic comparison	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Co-assembly	No	No	No	No	No	Yes	No	Yes	Yes
Binning	No	Support	No	No	No	Yes	No	Yes	Yes
Bin validation	No	Yes	No	No	No	No	No	No	Yes
Local Installation	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

sequencing experiments. By adequately setting all the pipeline's components, we were able to fully analyze completely individual metagenomes and even co-assemble related metagenomes using a desktop computer with only 16 GB RAM. The fully automatic nature of our system, not requiring any technical or bioinformatic knowledge, also makes it very easy to use. It is also completely independent of the availability of any Internet connection.

SqueezeMeta can be downloaded from https://github.com/jtamames/SqueezeMeta.

MATERIALS AND METHODS

SqueezeMeta is aimed to perform the analysis of several metagenomes in a single run. It can be run in three different modes (for a schematic workflow for the three modes, see **Figure 1**). These are:

1. Sequential mode: all metagenomes are treated individually and analyzed sequentially. This mode does not include binning, since each metagenome is treated independently.
2. Co-assembly mode: reads from all samples are pooled and a single assembly is performed. Reads from individual samples are then mapped back to the co-assembly, which enables obtaining the coverage of contigs and individual genes in these contigs. Based on these abundances, subsequent binning methods allow classifying contigs in genomic bins.
3. Merged mode: co-assembly is a very intensive process that requires plenty of computational resources, especially RAM. If the number of samples is high, requirements can easily exceed the capabilities of the computing infrastructure. SqueezeMeta's merged mode permit co-assembly of a large number of samples, using a procedure similar to the one used in the analysis of TARA Oceans metagenomes (Tully et al., 2018). Samples are first assembled individually. The resulting sets of contigs are merged by combining contigs with ≥99% semi-global identity, using CD-HIT (Fu et al., 2012). Then the remaining contigs are re-assembled using Minimus2 (Treangen et al., 2011) with parameters -D OVERLAP = 100 MINID = 95, to look for overlapping contigs coming from pieces of the same genome in different samples. The merging produces a single set of contigs, and the analysis proceeds as in the co-assembly mode.

SqueezeMeta uses a combination of custom scripts and external software packages for the different steps of the analysis. A more detailed description of these steps follows:

Data Preparation

A SqueezeMeta run only requires a configuration file indicating the metagenomic samples and the location of their corresponding sequence files. The program creates the appropriate directories and prepares the data for further steps.

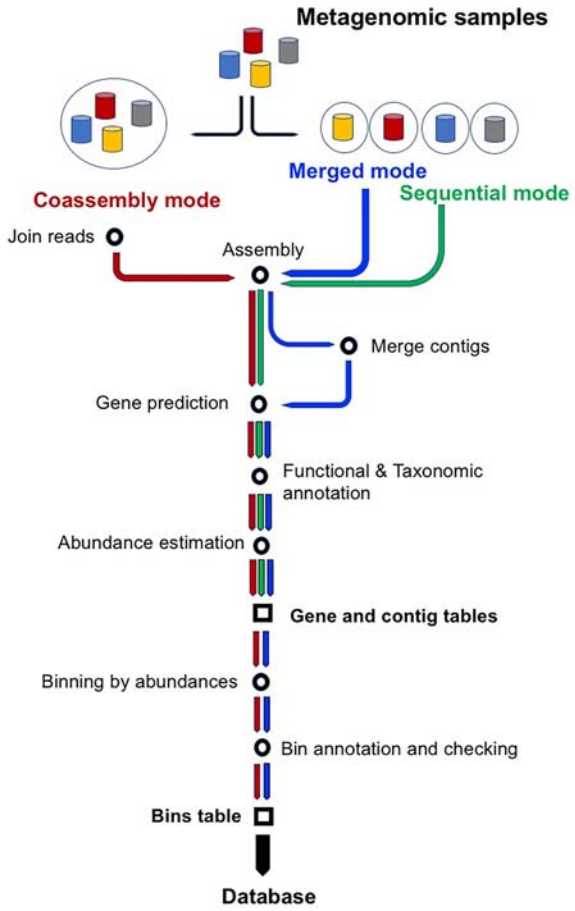


FIGURE 1 | Workflow of the three modes of operation of SqueezeMeta: sequential, co-assembly and merged. Starting from metagenomic samples, green, blue and red arrows indicate main steps in sequential, merged and co-assembly modes. All modes create two of the three main results tables: ORF and contig tables. Co-assembly and merged modes also apply binning and, therefore, they also create the bin table.

Trimming and Filtering

SqueezeMeta uses Trimmomatic for adapter removal, trimming and filtering by quality, according to the parameters set by the user (Bolger et al., 2014).

Assembly

When assembling large metagenomic datasets, computing resources, especially memory usage, are critical. SqueezeMeta uses Megahit (Li et al., 2015) as its reference assembler, since we find it has an optimal balance between performance and memory usage. SPAdes (Bankevich et al., 2012) is also supported. For assembly of the long, error-prone MinION reads, we use Canu (Koren et al., 2017). The user can select any of these assemblers. In the merged mode, each metagenome will be assembled separately and the resulting contigs will be merged and joined as outlined above. Either way, the resulting set of contigs is filtered by length using prinseq (Schmieder and Edwards, 2011), to discard short contigs if required.

Gene and rRNA Prediction

This step uses the Prodigal gene prediction software (Hyatt et al., 2010) to perform a gene prediction on the contigs, retrieving the corresponding amino acid sequences, and looks for rRNAs using barrnap (Seemann, 2014). The resulting 16S rRNA sequences are classified using the RDP classifier (Wang et al., 2007).

Homology Searching

SqueezeMeta uses the Diamond software (Buchfink et al., 2015) for comparison of gene sequences against several taxonomic and functional databases, because of its optimal computation speed while maintaining sensitivity. Currently, three different Diamond runs are performed: against the GenBank nr database for taxonomic assignment, against the eggNOG database (Huerta-Cepas et al., 2016) for COG/NOG annotation, and against the latest publicly available version of KEGG database (Kanehisa and Goto, 2000) for KEGG ID annotation. SqueezeMeta also classifies genes against the PFAM database (Finn et al., 2014), using HMMER3 (Eddy, 2009). These databases are installed locally and updated at the user's request.

Taxonomic Assignment of Genes

Custom scripts are used for this step of the analysis. For taxonomic assignment, SqueezeMeta implements a fast LCA algorithm that looks for the last common ancestor of the hits for each query gene using the results of the Diamond search against GenBank nr database (the most complete reference database available). For each query sequence, we select a range of hits having at least 80% of the bit-score of the best hit and differing by less than 10% of its identity percentage. The LCA is the lower rank taxon common to most hits, since a small number of hits belonging to other taxa are allowed to add resilience against, for instance, annotation errors. Importantly, our algorithm includes strict cut-off identity values for the various taxonomic ranks. This means that hits must pass a minimum amino acid identity level to be used for assigning to a particular taxonomic rank. These thresholds are 85, 60, 55, 50, 46, 42, and 40% for species, genus, family, order, class, phylum, and superkingdom ranks, respectively (Luo et al., 2014). Hits below these identity levels cannot be used to make assignments to the corresponding rank. For instance, a protein will not be assigned to species level if it has no hits above 85% identity. Moreover, a protein will remain unclassified if it has no hits above 40% identity. Inclusion of these thresholds guarantees that no assignments are performed based on weak, inconclusive hits.

Functional Assignments

Genes in COGs and KEGG IDs can be annotated using the classical best hit approach or a more sensitive one considering the consistency of all hits (**Supplementary Methods** in **Supplementary File S1**). In short, the first hits exceeding an identity threshold for each COG or KEGG are selected. Their bitscores are averaged, and the ORF is assigned to the highest-scoring COG or KEGG whose score exceeds the score of any other by 20%, otherwise the gene remains unannotated.

This procedure does not annotate conflicting genes with close similarities to more than one protein family.

Taxonomic Assignment of Contigs and Disparity Check

The taxonomic assignments of individual genes are used to produce consensus assignments for the contigs. A contig is annotated to the taxon to which most of their genes belong (**Supplementary File S1**). The required percentage of genes assigned to that taxon can be set by the user, so that it is possible to accommodate missing or incorrect annotations of a few genes, recent HGT events, etc. A disparity score is computed for each contig, indicating how many genes do not concur with the consensus (**Supplementary File S1**). Contigs with high disparity could be flagged to be excluded from subsequent analyses.

Coverage and Abundance Estimation for Genes and Contigs

To estimate the abundance of each gene and each contig in each sample, SqueezeMeta relies on mapping of original reads onto the contigs resulting from the assembly. The software Bowtie2 (Langmead and Salzberg, 2012) is used for this task, but we also included Minimap2 (Li, 2018) for mapping long MinION reads. This is followed by Bedtools (Quinlan and Hall, 2010) for extraction of the raw number of reads and bases mapping to each gene and contig. Custom scripts are used to compute the average coverage and normalized RPKM values that provide information on gene and contig abundance.

In sequential mode, SqueezeMeta would stop here. Any of the co-assembly modes allow binning the contigs for delineating genomes.

Binning

Using the previously obtained contig coverage in different samples, SqueezeMeta uses different binning methods to separate contigs putatively coming from the same organism. Basically, binning algorithms classify contigs coming from the same genomes because their coverages covary along the samples, and their oligonucleotide composition is similar. Currently, Maxbin (Wu et al., 2015) and Metabat2 (Kang et al., 2015) are supported. In addition, SqueezeMeta includes DAS Tool (Sieber et al., 2018) to merge the multiple binning results in just one set.

SqueezeMeta calculates average coverage and RPKM values for the bins in the same way as above, mapping reads to the contigs belonging to the bin.

Taxonomic Assignment of Bins and Consistency Check

SqueezeMeta generates a consensus taxonomic assignment for the bins in the same way as it did for the contigs. A bin is annotated to the consensus taxon, that is, the taxon to which most of its contigs belong. As previously, a disparity score is computed for each bin, indicating how many of the contigs are discordant with the bin's consensus taxonomic assignment. This can be used as an initial measure of the bin's possible contamination.

Bin Check

The goodness of the bins is estimated using the CheckM software (Parks et al., 2015). In short, CheckM provides indications of a bin's completeness, contamination and strain heterogeneity by creating a profile of single-copy, conserved genes for the given taxon and evaluating how many of these genes were found (completeness), and how many were single-copy (contamination and strain heterogeneity). SqueezeMeta automates CheckM runs for each bin, using the consensus annotation for the bin as the suggested taxonomic origin.

Merging of Results

Finally, the system merges all these results and generates several tables: (1) a gene table, with all the information regarding genes (taxonomy, function, contig and bin origin, abundance in samples, and amino acid sequence). (2) A contig table, gathering all data for the contigs (taxonomy, bin affiliation, abundance in samples, and disparity), and (3) A bin table with all information related to the bins (taxonomy, completeness, contamination, abundance in samples, and disparity).

Database Creation

These three tables and the optional metadata will be used to create a MySQL database for easy inspection of the data arising from the analysis. The database includes a web-based user interface that enables easy creation of queries, so that the user does not need to have any knowledge on database usage to operate it (**Figure 2**). The interface allows queries on one table (genes, contigs or bins) or combinations of tables, enabling complex questions such as "Retrieve contigs having genes related to trehalose from Bacteroidetes more abundant than 5x coverage in sample X" or "Retrieve antibiotic resistance genes active in one condition but not in another". The resulting information can be exported to a table.

When combining metagenomes and metatrancriptomes, the latter can be analyzed in a straightforward way by just mapping the cDNA reads against the reference metagenomes. In this way, we can obtain and compare the abundances of the same genes in both the metagenome and the metatranscriptome. However, this will obviate these genes present only in the latter, for instance genes belonging to rare species in the metagenome (therefore unassembled) and that happen to be very active. SqueezeMeta can deal with this situation using the merged mode. Metagenomes and metatranscriptomes are assembled separately and then merged so that contigs can come from DNA from the metagenome, cDNA from the metatranscriptome or both. Normalization of read counts makes it possible to compare presence and expression values within or between different samples.

RESULTS

To illustrate the use of the SqueezeMeta software, we analyzed 32 metagenomic samples corresponding to gut microbiomes of Hadza and Italian subjects (Rampelli et al., 2015), using the three modes of analysis. The total number of reads for all metagenomes

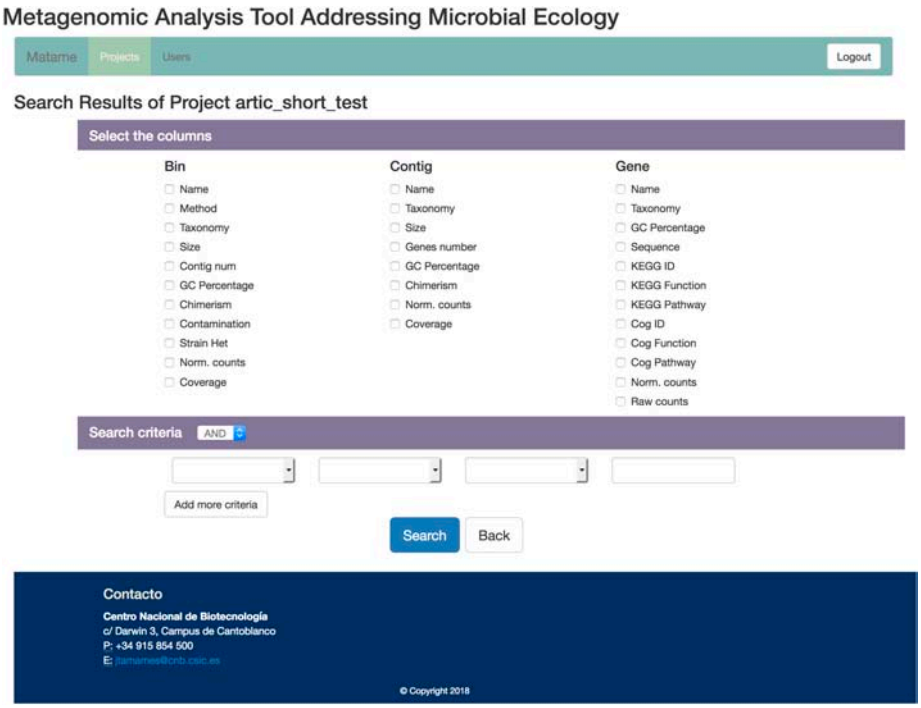


FIGURE 2 | Snapshot of the SqueezeMeta user interface to its database. A flexible and intuitive system for building queries allows interrogating the database with complex questions involving combination of data from different tables.

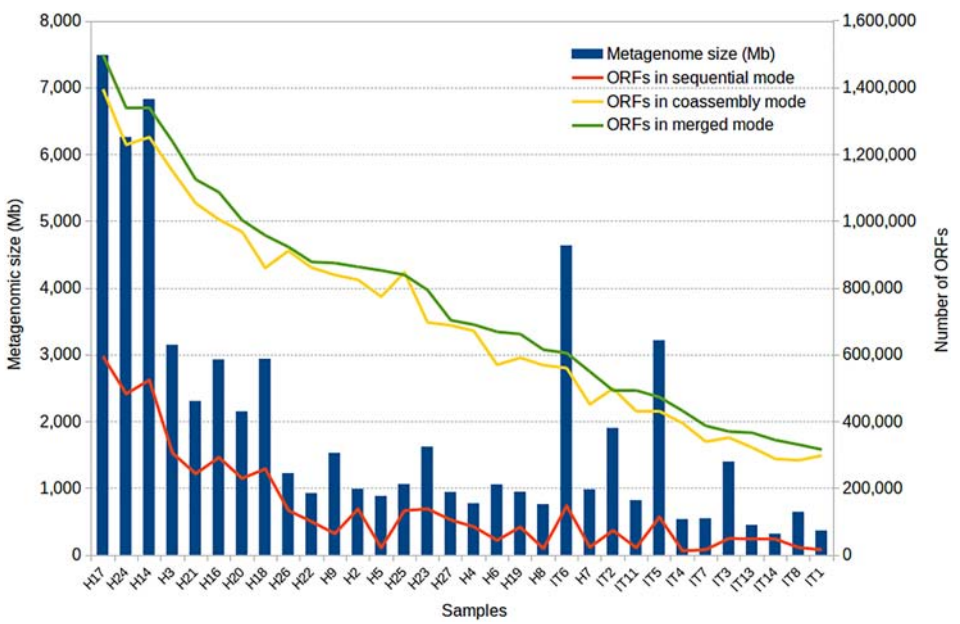


FIGURE 3 | Results of the application of SqueezeMeta to 32 gut metagenomes of Hadza (H) and Italian (IT) subjects. The figure shows the size of the metagenomes and the number of genes obtained by the three modes of analysis.

is 829.163.742. We used a 64-CPU computer cluster with 756 GB RAM in the National Center for Biotechnology, Madrid, Spain. After discarding contigs below 200 bps, the total number of genes was 4,613,697, 2,401,848, and 2,230,717 for the sequential, merged and co-assembled modes, respectively. Notice that the number of genes is lower in the two latter modes that involve co-assembly since the genes present in more than one metagenome will be counted just once in the co-assembly (they are represented

by just one contig product of the co-assembly) but more than once in the individual samples (they are present in one different contig per sample). A more accurate comparison is shown in **Figure 3**, where a gene in the co-assembly is assumed to be present in a given sample if it can recruit some reads from that sample. As co-assemblies create a much larger reference collection of contigs than individual metagenomes alone, even genes represented by a few reads in a sample can be identified by recruitment, while they will probably fail to assemble in the individual metagenome because of their low abundance. In other words, co-assembly will produce contigs and genes from abundant taxa in one or more samples, that can be used to identify the presence of the same genes in samples in which these original taxa are rare. Therefore, it enables discovering the presence of many more genes in each sample.

The improvement of gene recovery for the smaller samples is also noticeable by the percentage of mapped reads. The individual assembly for small samples achieves barely 35% of read mapping to the assembled metagenome, indicating that most reads could not be used. The small size (and therefore low coverage) of the metagenome prevented these reads from being assembled. When co-assembling these samples with the rest, more than 85% of the reads could then be mapped to the reference metagenome, indicating that co-assembly is able to capture most of the diversity found in these small samples.

Table 2 shows the characteristics of the analysis. Even if the merged mode obtains more contigs and genes than the co-assembly mode, we can see that the number of putatively inconsistent contigs (having genes annotated to different taxa) is lower in the second. Therefore, the co-assembly mode is more accurate than the merged mode, but the latter has the advantage of being able to work with an almost unlimited number of metagenomes because of its lower requirements.

Binning results have been analyzed according to the completeness and contamination values provided by CheckM (**Table 3**). Again, there are differences between the merged and the co-assembly modes, with the first providing more but less complete bins, and the latter giving bins of higher quality. Both modes are capable of obtaining quasi-complete genomes for tens of species, and hundreds of less complete genomes.

Figure 4 shows the abundance distribution of bins in samples. Italian subjects reveal a clear distinctive profile that make them cluster together. Bins belonging to the genera *Bacteoides* and *Faecalibacterium* are more abundant in these individuals than in Hadza individuals. The Hadza have increased diversity and fall into different groups corresponding to the presence of diverse species, in accordance with the distinctions found using functional profiles (Rampelli et al., 2015). The microbiota of these individuals contains genera such as *Allistipes* or *Prevotella* not present in the Italian metagenomes. Moreover, Spirochaetes from the genera *Treponema* are only present in Hadza subjects, which are supposedly not associated with pathogenesis. This information is directly retrieved from SqueezeMeta results and offers a revealing view of the genomic composition and differences between the samples. A similar result can be obtained for the functional annotations. The original functions represented in the bins can be used to infer the presence of metabolic pathways using the MinPath

TABLE 2 | Statistics on contigs and bins for the three SqueezeMeta modes on Hadza & Italian metagenomes.

	Merged mode	Co-assembly mode	Sequential mode
Number of contigs	893,438	983,350	2,478,560
N50	3900	2357	2854
Average percentage of mapped reads	85.01	89.47	74.47
Contigs with phylum annotation	719,098 (80.4%)	759,903 (77.2%)	1,951,445 (78.7%)
Contigs with disparity > 0	6626 (0.7%)	3772 (0.4%)	7588 (0.3%)
Highly inconsistent contigs (disparity > 0.25)	4496 (0.5%)	2993 (0.3%)	5433 (0.2%)
Number of genes	2,401,848	2,230,717	4,613,697
Genes with COG function	1,098,635 (45.7%)	982,029 (44.0%)	2,164,980 (46.9%)
Genes with KEGG function	835,498 (34.8%)	749,892 (33.6%)	1,683,636 (36.5%)
Total bins	563	423	N/A
Bins > 90% complete	120	115	N/A
Bins > 50% complete	359	192	N/A
High-quality bins (>90% complete, <10% contam)	50	67	N/A
Good quality bins (>75% complete, <10% contam)	82	112	N/A

Binning statistics refer to MaxBin results.

TABLE 3 | Example of some relevant high-quality bins (>90% completion, <10% contamination) obtained by the co-assembly mode of Hadza & Italian metagenomes.

Taxa	Size (bp)	Completeness	Contamination
o: Clostridiales	3,098,646	99.53%	3.57%
g: Bacteroides	3,521,779	99.45%	0.18%
o: Aeromonadales	2,579,625	99.16%	0.69%
g: Akkermansia	3,031,328	98.94%	4.31%
g: Treponema	2,879,091	98.62%	4.78%
g: Prevotella	3,354,102	97.51%	0.47%
s: *Escherichia coli*	4,710,119	97.49%	6.01%
g: Bifidobacterium	2,266,937	97.40%	3.70%
g: Megasphaera	2,490,127	96.93%	2.88%
s: *Succinatimonas sp.*	2,250,348	96.61%	4.22%
g: Parabacteroides	4,558,677	96.57%	3.18%
s: *Alistipes putredinis*	2,258,860	95.28%	5.77%
s: *Oscillibacter sp.*	1,801,182	95.11%	4.20%
s: *Bacteroides sp.*	3,901,726	95.08%	7.76%

Taxa are labeled according to their taxonomic rank. g, Genus; o, Order; s, Species.

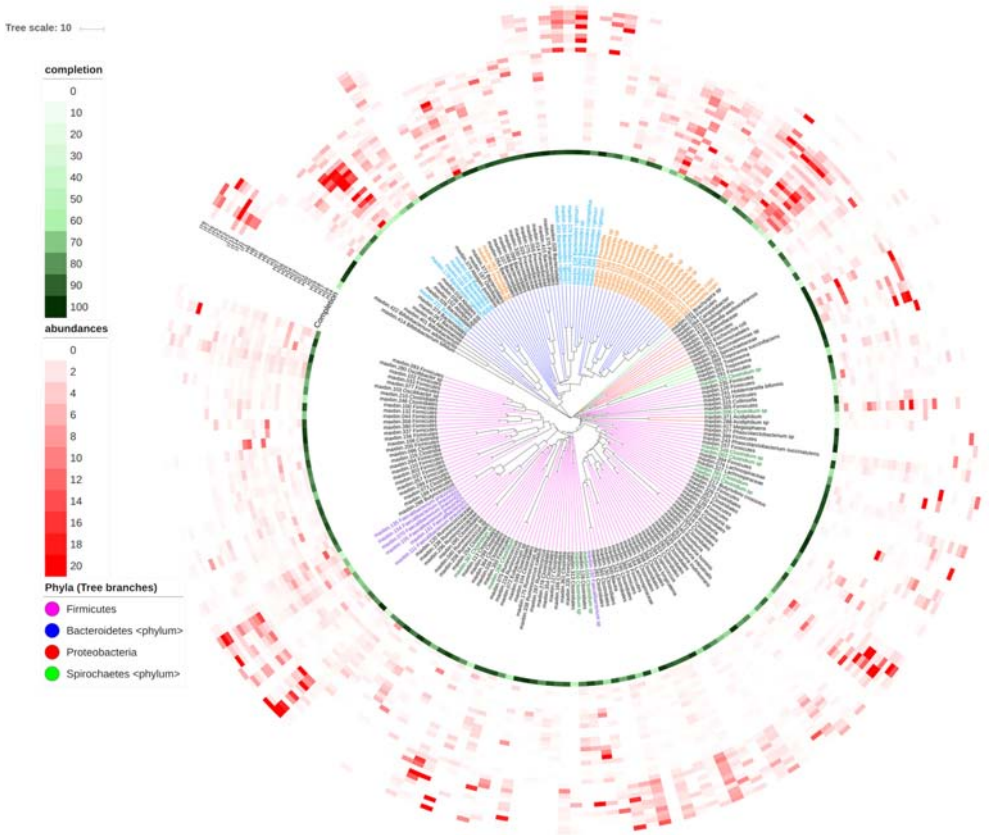


FIGURE 4 | Abundance of bins in the diverse samples. Bins were compared with the CompareM software (https://github.com/dparks1134/CompareM) to estimate their reciprocal similarities. The distances calculated between the bins were used to create a phylogenetic tree illustrating their relationships. The tree is shown in the inner part of the Figure. Branches in the tree corresponding to the four more abundant phyla in the tree (Firmicutes, Bacteriodetes, Proteobacteria, and Spirochaetes) were colored. Bins were named with their id number and original genera, and labels for the most abundant genera were also colored. Outer circles correspond to: the completeness of the bins (green-colored, most internal circle), and the abundance of each bin in each sample (red-colored). Each circle corresponds to a different sample (H, Hadza; I, Italians), and the red color intensities correspond to the bin's abundance in the sample. The picture was prepared using the iTOL software (https://itol.embl.de).

algorithm (Ye and Doak, 2009), that defines each pathway as an unstructured gene set and selects the fewest pathways that can account for the genes observed within each bin. The inference of several carbohydrate degradation pathways in the bins can be observed in **Supplementary Figure S1**.

One of the motivations for the development of SqueezeMeta was making it capable of performing a full metagenomic analysis on a limited computing infrastructure, such as the one that can be expected in the course of *in situ* metagenomic sequencing (Lim et al., 2014; Johnson et al., 2017). We created a setting mode (–lowmem) carefully tailored to run with limited amounts of resources, especially RAM memory. To test this capability, we were able to co-assemble two metagenomic samples from the Hadza metagenomes, composed of 40 million reads amounting to almost 4 GB of DNA sequence. We ran the merged mode of SqueezeMeta using the – low-memory option in a standard laptop computer, using just 8 cores and 16 GB RAM. The run was completed in 10 h, generating 33,660 contigs in 38 bins and 124,065 functionally and taxonomically annotated genes. Using the same settings, we also co-assembled ten MinION metagenomes from the gut microbiome sequencing of head and neck cancer patients[1], summing 581 MB in less than 4 h. These experiments reveal that SqueezeMeta can be run even with scarce computational resources, and it is suitable for its intended use of *in situ* sequencing where the metagenomes will be moderate in size.

DISCUSSION

SqueezeMeta is a highly versatile pipeline that enables analyzing a large number of metagenomes or metatranscriptomes in a very straightforward way. All analysis steps are included, starting with assembly, subsequent taxonomic/functional assignment of the resulting genes, abundance estimation

[1]https://www.ncbi.nlm.nih.gov/bioproject/PRJNA493153

and binning to obtain as many genomes as possible in the samples. SqueezeMeta is designed to run in moderately-sized computational infrastructures, relieving the burden of co-assembling tens of metagenomes by using sequential metagenomic assembly and ulterior merging of resulting contigs. The software includes specific software and adjustments to be able to process MinION sequences.

The program includes several verifications on the results, such as the detection of possible inconsistent contigs and bins, and estimation of the latter's completion using the checkM software. Finally, results can easily be inspected and managed since SqueezeMeta includes a built-in MySQL database that can be queried via a web-based interface, allowing the creation of complex queries in a very simple way.

One of the most remarkable features of this software is its capability to operate in limited computing infrastructure. We were able to analyze several metagenomes in a few hours using a virtual machine with just 16 GB RAM. Therefore, SqueezeMeta is apt to be used in scenarios in which computing resources are limited, such as remote locations in the course of metagenomic sampling campaigns. Also, it does not require the availability of any Internet connection. Obviously, complex, sizeable metagenomes cannot be analyzed with these limited resources. However, the intended use of *in situ* sequencing will likely produce a moderate and manageable data size.

SqueezeMeta will be further expanded by the creation of new tools allowing in-depth analyses of the functions and metabolic pathways represented in the samples.

AUTHOR CONTRIBUTIONS

JT conceived and designed the tool and wrote the manuscript. JT and FP-S created the software and performed all necessary testing and read and approved the manuscript.

FUNDING

This research was funded by projects CTM2016-80095-C2-1-R and CTM2013-48292-C3-2-R, Spanish Ministry of Economy and Competitiveness. This manuscript was made available as a pre-print at BioRxiv (Tamames and Puente-Sanchez, 2018).

ACKNOWLEDGMENTS

We are grateful to Natalia García-García for helping to test the system.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.03349/full#supplementary-material

FIGURE S1 | The presence of several carbohydrate degradation pathways in the bins. The outer circles indicate the percentage of genes from a pathway present in each of the bins. According to that gene profile, MinPath estimates whether or not the pathway is present. Only pathways inferred to be present are colored. As in **Figure 4**, the bins tree is performed from a distance matrix of the orthologous genes' amino acid identity, using the compareM software (https://github.com/dparks1134/CompareM). The four most abundant phyla are colored (branches in the tree), as well as the most abundant genera (bin labels). The picture was elaborated using the iTOL software (https://itol.embl.de).

FILE S1 | Description of novel algorithms implemented in SqueezeMeta.

REFERENCES

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.* 8:2358. doi: 10.1371/journal.pcbi.1002358

Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J., and Bork, P. (2010). SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* 26, 2977–2978. doi: 10.1093/bioinformatics/btq536

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

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

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Carr, R., and Borenstein, E. (2014). Comparative analysis of functional metagenomic annotation and the mappability of short reads. *PLoS One* 9:e105776. doi: 10.1371/journal.pone.0105776

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524. doi: 10.1038/nbt.3423

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform.* 23, 205–211. doi: 10.1142/9781848165632_0019

Eren, A. M., Esen, C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., et al. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3, 1–29. doi: 10.7717/peerj.1319

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230.

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

Glass, E. M., and Meyer, F. (2011). "The metagenomics RAST server: a public resource for the automatic phylogenetic and functional analysis of metagenomes," in *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, (Hoboken, NJ: Wiley Blackwell), 325–331. doi: 10.1002/9781118010518.ch37

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

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). EGGNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293. doi: 10.1093/nar/gkv1248

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y., and Tighe, S. W. (2017). Real-time DNA sequencing in the antarctic dry valleys using the Oxford nanopore sequencer. *J. Biomol. Tech.* 28, 2–7. doi: 10.7171/jbt.17-2801-009

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopaedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. doi: 10.7717/peerj.1165

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., and Zhan, X. (2016). FMAP: functional mapping and analysis pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics* 17:420. doi: 10.1186/s12859-016-1278-0

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: 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. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 10:359. doi: 10.1186/1471-2105-10-359

Lim, Y. W., Cuevas, D. A., Silva, G. G. Z., Aguinaldo, K., Dinsdale, E. A., Haas, A. F., et al. (2014). Sequencing at sea: challenges and experiences in Ion Torrent PGM sequencing during the 2013 Southern Line Islands Research Expedition. *PeerJ* 2:e520. doi: 10.7717/peerj.520

Luo, C., Rodriguez-R, L. M., and Konstantinidis, K. T. (2014). MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* 42:e73. doi: 10.1093/nar/gku169

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., and Kubal, M. (2008). The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Narayanasamy, S., Jarosz, Y., Muller, E. E. L., Heintz-Buschart, A., Herold, M., Kaysen, A., et al. (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* 17:260. doi: 10.1186/s13059-016-1116-8

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Pignatelli, M., Aparicio, G., Blanquer, I., Hernández, V., Moya, A., and Tamames, J. (2008). Metagenomics reveals our incomplete knowledge of global diversity. *Bioinformatics* 24, 2124–2125. doi: 10.1093/bioinformatics/btn355

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol.* 16:114. doi: 10.1186/s13059-015-0677-2

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232. doi: 10.1038/nature16996

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rampelli, S., Schnorr, S. L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., et al. (2015). Metagenome sequencing of the hadza hunter-gatherer gut microbiota. *Curr. Biol.* 25, 1682–1693. doi: 10.1016/j.cub.2015.04.055

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., et al. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3, 836–843. doi: 10.1038/s41564-018-0171-1

Tamames, J., and Puente-Sanchez, F. (2018). SqueezeM, a fully automatic metagenomic analysis pipeline from reads to bins. *bioRxiv* [Preprint]. doi: 10.1101/347559

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., and Pop, M. (2011). Next generation sequence assembly with AMOS. *Curr. Protoc.* 33, 11.8.1–11.8.18. doi: 10.1002/0471250953.bi1108s33

Tully, B. J., Graham, E. D., and Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* 5:170203. doi: 10.1038/sdata.2017.203

Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP - a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6:158. doi: 10.1186/s40168-018-0541-1

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naïve 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

Westreich, S. T., Treiber, M. L., Mills, D. A., Korf, I., and Lemay, D. G. (2018). SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinformatics* 19:175. doi: 10.1186/s12859-018-2189-z

Wommack, K. E., Bhavsar, J., and Ravel, J. (2008). Metagenomics: read length matters. *Appl. Environ. Microbiol.* 74, 1453–1463. doi: 10.1128/AEM.02181-07

Wu, Y. W., Simmons, B. A., and Singer, S. W. (2015). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi: 10.1093/bioinformatics/btv638

Ye, Y., and Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.* 5:465. doi: 10.1371/journal.pcbi.1000465

Conflict of Interest Statement: 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.

Copyright © 2019 Tamames and Puente-Sánchez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 29 January 2019
doi: 10.3389/fmicb.2019.00052

Exploring the Fecal Microbial Composition and Metagenomic Functional Capacities Associated With Feed Efficiency in Commercial DLY Pigs

Jianping Quan[1†], Gengyuan Cai[1,2†], Ming Yang[2], Zhonghua Zeng[1], Rongrong Ding[1], Xingwang Wang[1], Zhanwei Zhuang[1], Shenping Zhou[1], Shaoyun Li[1], Huaqiang Yang[1], Zicong Li[1], Enqin Zheng[1], Wen Huang[3], Jie Yang[1] and Zhenfang Wu[1]**

[1] College of Animal Science and National Engineering Research Center for Breeding Swine Industry, South China Agricultural University, Guangzhou, China, [2] National Engineering Research Center for Breeding Swine Industry, Guangdong Wens Foodstuffs Group Co., Ltd., Guangzhou, China, [3] Department of Animal Science, Michigan State University, East Lansing, MI, United States

OPEN ACCESS

Edited by:
Xing Chen,
China University of Mining and Technology, China

Reviewed by:
Xianwen Ren,
Peking University, China
Robert Heyer,
Otto-von-Guericke-Universität Magdeburg, Germany

***Correspondence:**
Jie Yang
jieyang2012@hotmail.com
Zhenfang Wu
wzfemail@163.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: 22 October 2018
Accepted: 14 January 2019
Published: 29 January 2019

Citation:
Quan J, Cai G, Yang M, Zeng Z, Ding R, Wang X, Zhuang Z, Zhou S, Li S, Yang H, Li Z, Zheng E, Huang W, Yang J and Wu Z (2019) Exploring the Fecal Microbial Composition and Metagenomic Functional Capacities Associated With Feed Efficiency in Commercial DLY Pigs. Front. Microbiol. 10:52. doi: 10.3389/fmicb.2019.00052

Gut microbiota has indispensable roles in nutrient digestion and energy harvesting, especially in processing the indigestible components of dietary polysaccharides. Searching for the microbial taxa and functional capacity of the gut microbiome associated with feed efficiency (FE) can provide important knowledge to increase profitability and sustainability of the swine industry. In the current study, we performed a comparative analysis of the fecal microbiota in 50 commercial Duroc × (Landrace × Yorkshire) (DLY) pigs with polarizing FE using 16S rRNA gene sequencing and shotgun metagenomic sequencing. There was a different microbial community structure in the fecal microbiota of pigs with different FE. Random forest analysis identified 24 operational taxonomic units (OTUs) as potential biomarkers to improve swine FE. Multiple comparison analysis detected 8 OTUs with a significant difference or tendency toward a difference between high- and low-FE pigs ($P < 0.01$, $q < 0.1$). The high-FE pigs had a greater abundance of OTUs that were from the Lachnospiraceae and Prevotellaceae families and the *Escherichia-Shigella* and *Streptococcus* genera than low-FE pigs. A sub-species *Streptococcus gallolyticus* subsp. *gallolyticus* could be an important candidate for improving FE. The functional capacity analysis found 18 KEGG pathways and CAZy EC activities that were different between high- and low-FE pigs. The fecal microbiota in high FE pigs have greater functional capacity to degrade dietary cellulose, polysaccharides, and protein and may have a greater abundance of microbes that can promote intestinal health. These results provided insights for improving porcine FE through modulating the gut microbiome.

Keywords: DLY pigs, feed efficiency, gut microbiota, 16S rRNA gene, metagenome sequencing

INTRODUCTION

Feed cost accounts for nearly 70% of the total cost in pig production (Teagasc, 2015). Therefore, improving feed efficiency (FE) of the pig will reduce feeding expense and increase profitability while also reducing the environmental impact of pig production (Rotz, 2004). In the commercial pig population, especially in Duroc × (Landrace × Yorkshire) (DLY) pigs, the improvement in FE will bring obvious benefits. The FE can be measured by using residual feed intake (RFI) or the feed conversion ratio (FCR). The FCR is calculated as the feed intake divided by the body weight gained. In other words, the high-FCR individuals are less efficient at converting feed into body weight than the low-FCR individuals. The FE in this study was measured by FCR.

In recent years, analyzing the microbiota of breeding animals has gained interest because it allows for the prediction of the potential function and associated metabolites of such communities, which are believed to impact all aspects of host physiology including nutrient processing, energy harvesting, and animal performance (Hiergeist et al., 2015; Xiao et al., 2016; Ferrario et al., 2017; Fouhse et al., 2017; Tan et al., 2017). Previous studies have revealed a possible link between the intestinal microbiota and FE in pigs; e.g., Tan et al. (2018) discovered that in Landrace pigs, the high-FCR pigs had a greater abundance of *Lactobacillus* and *Streptococcus* than the low-FCR pigs. In Large White × Landrace pigs, there was a greater abundance of *Christensenellaceae*, *Oscillibacter*, and *Cellulosilyticum* in the gastrointestinal tract of high-FE pigs (McCormack et al., 2017). In Duroc pigs, Yang et al. (2016) identified 31 operational taxonomic units (OTUs) showing potential associations with FE. Interestingly, these studies also imply that in different breeds of pigs, there may be differences in microbial composition and advantage species. Xiao et al. (2017) also indicated that breed-specific bacteria in swine intestinal tract may exist, even when pigs were treated with the same diet, farm conditions, and management methods.

There are few studies that focus on the association between microbial composition and functional capacity in regards to FE in DLY pigs. For DLY pigs, which are the largest population in the world porcine industry, understanding the relationships between the intestinal tract and host FE performance is meaningful. In our previous studies, we found that DLY pigs with contrasting FE have 11, 55, and 55 OTUs that were different among ileum, cecum and colon (Quan et al., 2018). The functional predictive analysis suggested that the microbial fermentation in cecum and colon may play important roles in improving porcine FE. However, due to the limitations of the research strategy, we have not been able to annotate the microbial gene into more functional database and get more detailed microbial classification differences between high- and low- FE pigs. In this study, we used 16S rRNA gene sequencing and high-throughput metagenomic sequencing to investigate whether the microbiota composition and potential functionality of the intestinal microbiota are linked with FE.

MATERIALS AND METHODS

Animals and Sample Collection

This study was conducted according to the protocols approved by the Animal Care and Use Committee (ACUC) of the South China Agricultural University (SCAU) (approval number SCAU#0017). In an experimental pig farm (Guangdong, Yunfu, Southern China), a total of 226 normal weaning (28-day-old) commercial DLY female pigs were randomly raised in a fattening house comprised of 30 pens, each housing 6–8 pigs. All of the pigs that were analyzed in this study were selected from populations with similar genetic backgrounds and were the same gender. During the fattening stage, the pigs were raised with the same customized diet in man-controlled farm conditions and similar management conditions. The customized corn-soybean feed (free of probiotics and antibiotics) contained 16% crude protein, 3100 kJ of digestible energy and 0.78% lysine. The diet was available *ad libitum* from an automatic feeding trough, Osborne's FIRE (Feed Intake Recording Equipment) System (Osborne Industries inc, Osborne, Kansas), which can separately record daily feed intake and daily body weight gain of each pig. Water was available *ad libitum* from nipple drinkers. During the whole experiment, any pigs treated with antibiotics were removed from the study. The FCR values of all pigs were calculated at 140 days of age. After the FCR value ranking of each pig, the 25 pigs with the lowest FCR (highest FE) and the 25 pigs with the highest FCR (lowest FE) were selected for this study. The fecal samples of 50 sows were collected following rectal stimulation and were transferred immediately to liquid nitrogen for temporary storage. Then, the samples were sent to the laboratory where they were stored at −80°C until analysis. We further chose six fecal samples for metagenomic sequencing. These six pigs included three individuals from the high-FE group and their full siblings from the low-FE group.

DNA Extraction, PCR Amplification, and 16S rRNA Gene Sequencing

Fecal DNA was extracted using a Soil Genome^{TM} DNA Isolation Kit (Qiagen, Düsseldorf, Germany) in accordance with the manufacturer's instructions. DNA concentration and quality were measured using UV-Vis spectrophotometry (NanoDrop 2000, Waltham, MA, United States) and agarose gel electrophoresis. The DNA obtained from each sample was diluted to 1 ng/μL with sterile water. Amplification of the V4–V5 hypervariable region of the bacterial 16S rRNA gene was performed using universal primers, where the reverse primer contained a 6-bp error-correcting barcode unique to each sample (515f: 5′-GTGCCAGCMGCCGCGGTAA-3′, 907r: 5′-CCGTCAATTCCTTTGAGTTT-3′). Amplification was performed using an initial denaturation at 98°C for 1 min followed by 30 cycles of denaturation at 98°C for 10 s, annealing at 50°C for 30 s, elongation at 72°C for 30 s, and a final step at 72°C for 5 min. All PCR reactions were carried out using Phusion® High-Fidelity PCR Master Mix (NEB, Ipswich, MA, United States). PCR products were run in an electrophoresis chamber on a 2% agarose gel to confirm the successful

amplification of the target gene. DNA bands of 400–450 bp, corresponding to the 16S rRNA gene amplicon, were excised and purified using the GeneJET Gel Extraction Kit (Thermo Fisher Scientific, Waltham, MA, United States) according to the manufacturer's instructions. Purified amplicons were used for library preparation and pyrosequencing. Sequencing libraries were generated using NEB Next® Ultra™ DNA Library Prep Kit for Illumina (NEB, Ipswich, MA, United States), following the manufacturer's recommendations, and index codes were added. A Qubit@ 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, United States) and Agilent Bioanalyzer 2100 system were used to assess the quality of the library. Pyrosequencing was performed on the Illumina HiSeq 2 × 250 platform (Illumina, San Diego, CA, United States). The 16S rRNA gene sequence data have been deposited in the NCBI SRA database with an accession number of SUB4418365.

Processing of Sequencing Data

Sequencing reads were assigned to each sample, based on unique barcodes, and truncated by cutting off the barcode and primer sequence. The original DNA fragments were merged into tags using FLASH (v1.2.7) (Magoc and Salzberg, 2011). Quality filtering of the raw tags was performed under specific filtering conditions to generate high-quality clean tags according to the QIIME (v1.9.1) quality-controlled process (Caporaso et al., 2010). To generate effective tags, the chimeric sequences were removed from clean tags using the UCHIME algorithm based on the reference database (Gold database) (Haas et al., 2011). After selecting representative species for each OTU, each of the remaining sequences was assigned to an OTU when at least 97% threshold identity was obtained using UPARSE software (v7.0.1) (Edgar, 2013). The taxonomy of each OTU representative sequence was assigned for further annotation using the RDP Classifier algorithm[1] (Wang et al., 2007) against the SILVA ribosomal RNA gene database. Subsequent analyses were performed based on the OTU information. A Venn diagram was generated using the *VennDiagram* R package to show shared and unique OTUs between high- and low-FE pigs.

In this study, we used mothur software (v.1.30.1) to calculate the community alpha diversity indices, including Chao1 and ACE indices, which estimate community richness, and Shannon and Simpson indices, which estimate community diversity (Schloss et al., 2011). A significant difference in alpha diversity between high- and low-FE groups was determined using the Mann–Whitney *U*-test. Moreover, we also calculated the community pan-OTU number and Good's coverage index to evaluate sample size and the sequencing depth. Principal component analysis (PCA) was determined to evaluate the community structure similarity between the samples in the high- and low-FE groups. Significant differences in beta-diversity across opposite FE groups were evaluated using permutational multivariate analysis of variance (PERMANOVA) with 10^4 permutations. In addition, the effects of pen information, initial weight and final weight on variance of sample microbial community composition were evaluated by PERMANOVA analyses (Anderson, 2001; Anderson and Walsh, 2013). Bacterial taxonomic distributions of sample communities were visualized using the *ggplot2* R package. In subsequent analyses, taxa occurring in less than three samples with a relative abundance less than 0.01% of the total community were removed. To test whether microbial community composition can predict feed conversion, we trained a random forest model at the OTU level on all samples based on a random sampling with replacement (Number of decision trees = 500). We evaluated the performance using 10-fold cross-validation. The cross-validation error curve (average of 5 test sets each) of the 10-fold cross-validation was averaged. The variable importance by mean decrease in accuracy was calculated. The predictive power was scored in a receiver operating characteristic (ROC) analysis. The discriminatory power of OTUs was calculated as the area under the ROC curve (AUC) using the *plotROC* R package.

The comparison of relative abundances of OTUs between high- and low-FE pigs was performed using Welch's *t*-test in STAMP software (White et al., 2009). The Benjamini–Hochberg False Discovery Rate (FDR) method (*q*-value) was used to correct the multiple comparisons (Benjamini and Hochberg, 1995). The statistical cutoff of the *p*-value <0.05 (Welch-Test) and *q*-value <0.05 (FDR) were set as the significance threshold. The relative abundance of different OTUs between high- and low-FE pigs was visualized by heatmap using *vegan* R package.

Metagenomic Sequencing and Statistical Analyses

Metagenome sequencing libraries were generated with an insert size of 350 base pairs (bp) for six fecal DNA samples following the manufacturer's instructions (Illumina, San Diego, CA, United States). The libraries for metagenomic analysis were sequenced on an Illumina HiSeq 2500 platform by an Illumina HiSeq – PE150 strategy. The raw reads were treated to remove reads with low qualities, trim the read sequences and remove adaptors using Readfq software (v8). The metagenomic sequencing data have been deposited in the NCBI SRA database with the accession number SUB4056369. Subsequently, pig genomic DNA sequences were removed by SOAPaligner software (v2.21) (Li et al., 2008).

De novo assembly of high quality reads was performed using SOAPdenove software (v2.04) with the parameters -d 1, -M 3, -R, -u, -F. Scaffolds were broken into new scaftigs at their gaps (Luo et al., 2012). Meanwhile, the scaftigs with a length less than 500 bp were removed, and the number of scaftigs $\geq$500 bp was calculated. The qualified scaftigs were applied to predict the bacterial open reading frames (ORFs) by MetaGeneMark (v2.10) software, and the sequences with lengths less than 100 bp were filtered out (Zhu et al., 2010). CD-HIT software (v4.5.8) was used to exclude the redundant genes from all predicted ORFs to construct a preliminary non-redundant gene catalog (Fu et al., 2012). Subsequently, clean reads of each sample were compared to the preliminary non-redundant gene catalog using SOAPaligner with the parameters of -m 200, -× 400, identity $\geq$95%. The number of reads was compared for each

[1] http://rdp.cme.msu.edu/

gene that could be calculated. The genes with a read number ≤2 were removed to obtain a final non-redundant gene catalog (Qin et al., 2012). The genes in the final non-redundant gene catalog were called unigenes. The abundance of a gene was calculated based on the number of reads that aligned to the gene, normalizing by the gene length and the total number of reads aligned to the unigenes (Karlsson et al., 2012). The specific formula for the relative abundance calculation of a gene was $G_k = \frac{r_k}{L_k} \cdot \frac{1}{\sum_{i=1}^{n} \frac{r_i}{L_i}}$., here r is the number of reads mapped to a gene and L is the length of gene. Subsequently, we used DIAMOND software (V0.7.9) to compare the unigenes with the Kyoto Encyclopedia of Genes and Genomes (KEGG) gene database to obtain KO annotation information and metabolic pathway information (Buchfink et al., 2015). We compared the unigenes with the Carbohydrate-Active enzymes database (CAZy) to obtain information on species and the functional classification of EC.

To determine the differential abundance of functional features between the high- and low-FE groups, Metastats analysis was applied (White et al., 2009). The Benjamini–Hochberg FDR method (q-value) was used to correct the multiple comparisons (Benjamini and Hochberg, 1995). Z-scores were calculated to construct a heatmap to demonstrate the relative abundance of the pathways in each group with the formula $z = (x-\mu)/\sigma$, where x is the relative abundance of the pathways in each group, μ is the mean value of the relative abundances of the pathways in all groups, and σ is the standard deviation of the relative abundances.

RESULTS

Phenotypic Values of Porcine FCR and Community Composition of Porcine Fecal Microbiota

All experimental pigs had daily feed intake and daily body weight gain separately recorded during the fattening stage (28-day-old to 140-day-old). The 25 pigs with the highest FE (FCR value: 2.29 ± 0.080) and the 25 pigs with the lowest FE (FCR value: 2.60 ± 0.088) were selected for this study. The FCR value was significantly different between the high- and low-FE groups (p-value < 0.001, **Figure 1A** and **Supplementary Table S1**).

A total of 50 pigs, which included extreme FCR values, were selected, and 16S rRNA gene sequencing was performed, which generated a total of 3,788,293 DNA sequence reads, aligned into 2,851,748 effective tags after quality control. Based on the 97% sequence similarity, the number of OTU samples ranged from 569 to 1037. The pan-OTU numbers of community would reach saturation when the sample size was greater than thirty (**Figure 1B**) and the Good's coverage indices in high- and low-FE groups were greater than 99% (**Figure 1C**), which indicated a sufficient sample size and adequate sequencing depth for

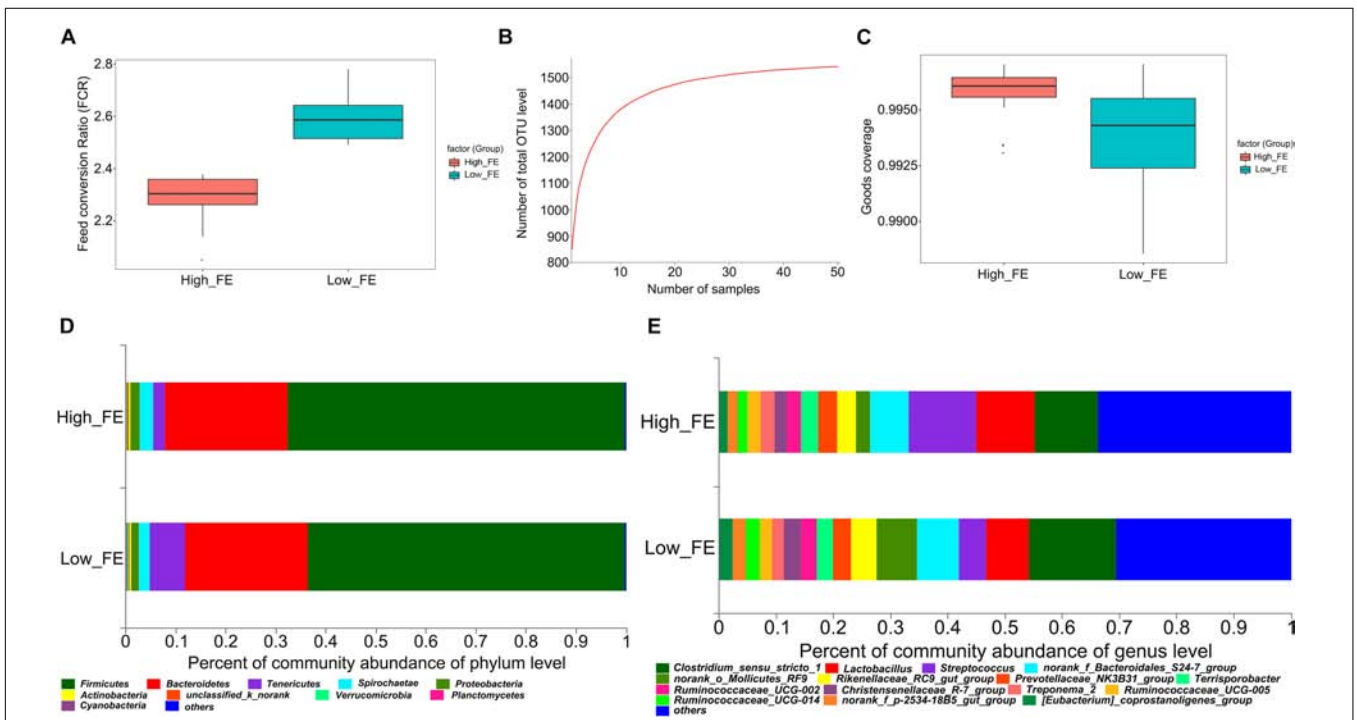


FIGURE 1 | The feed conversion ratio value (FCR), pan OTUs, Good's coverage and community composition in high- and low-feed-efficiency (FE) pigs. Groups are coded according to the feed efficiency status (High_FE, high feed efficiency; Low_FE, low feed efficiency). **(A)** FCR value in high- and low-FE pigs. **(B)** Pan OTU = sample size. The horizontal axis represents the number of samples. The vertical axis represents the number of OTUs contained in all samples. **(C)** Good's coverage value in high- and low-FE pigs. **(D)** Community composition at the phylum level in high- and low-FE pigs. **(E)** Community composition at the genus level in high- and low-FE pigs.

this study. The Venn diagrams show that 1437 OTUs were shared between the high- and low-FE groups. Only 44 and 60 OTUs were unique in the low- and high-FE groups, respectively (**Supplementary Figure S1A**).

These OTUs were annotated to the phylum, class, order, family and genus classification level. At the phylum level, the high- and low-FE pigs' microbial community was dominated by Firmicutes (67.47% vs. 63.35%), Bacteroidetes (24.40% vs. 24.68%), Tenericutes (2.43% vs. 7.01%), Spirochaetes (2.66% vs. 2.11%), and Proteobacteria (1.93% vs. 1.49%) (**Figure 1D**). At the genus level, *Streptococcus* (11.80%), *Clostridium sensu stricto 1* (11.11%), and *Lactobacillus* (10.19%) were the three most abundant genera in the high-FE group; *Clostridium sensu stricto 1* (15.21%), *Lactobacillus* (7.51%), and uncertain genera from Bacteroidales S24-7 (7.40%) were the three most abundant genera in the low-FE group (**Figure 1E**).

To further investigate microbial composition at the species level, shotgun metagenomic sequencing was performed in six fecal samples from three pairs of full-siblings having the high- and low-FCR phenotypes. The metagenomic sequencing produced a total of 56 Gbp of clean reads after removing low-quality sequences and host genomic DNA sequences. After subsequent assembly, a total of 1.15 million scaftigs with an average length of 1,095 bp and an average N50 length of 1,152 bp were produced. The phylogenetic composition of the fecal microbiota determined by shotgun metagenomic sequencing was similar to the result obtained in the 16S rRNA gene sequencing. Firmicutes, Bacteroidetes, Spirochaetes, and Proteobacteria were the dominant phyla (**Supplementary Figure S1B**). At the species level, we detected a total of 6,972 bacterial species from all six fecal samples. *Firmicutes bacterium CAG:110*, *Treponema bryantii* and *Bacteroides* sp. *CAG:1060* were the three most abundant species (**Supplementary Figure S1C**).

Comparison of Fecal Microbial Community Diversity Between High- and Low-FE Pigs

To evaluate the alpha-diversity of bacterial communities in high- and low-FE pigs, we compared the community richness indices (Chao1 and ACE) and diversity indices (Shannon and Simpson) of the microbiota in high- and low-FE pigs. We found that high-FE pigs have significantly higher Chao1 and ACE indices than low-FE pigs ($P < 0.01$, **Figure 2A** and **Supplementary Figure S2A**). However, the Shannon and Simpson indices were not significantly different between these two groups (**Figure 2B** and **Supplementary Figure S2B**). Based on the abundance profiling of the OTU level, PCA analysis showed that most of the samples could be clustered into two groups, which was very consistent with the grouping results according to performance of feed conversion (**Figure 2C**). A significant dissimilarity in beta-diversity between high- and low-FE groups was observed (PERMANOVA, p-value <0.01). Based on the abundance profiling of species level generated by metagenomic sequencing, there were also a clear difference in bacterial composition in the high- and low-FE pigs (**Supplementary Figure S2C**). In addition, we found that the initial weight and final weight had no significant effect on porcine fecal microbial composition (p-value >0.3). The pig pen had a tendency to make effect on microbial composition, but also cannot reach the significant level in our study (p-value = 0.077) (**Supplementary Table S2**).

Identification of Potential Biomarkers That Could Account for the FE Differences

To determine whether OTUs could serve as biomarkers to classify pigs into high- and low-FE groups and which OTUs play important roles in this process, we constructed a random forest model. The OTU-level random forest model had an error of 0.025 when the number of top important variables was 24 (**Figure 3A**). The mean decrease in accuracy of the top 24 important variables is shown in **Figure 3B**. Six OTUs that included the top three important variables (OTU509, OTU1013, and OTU197) for predicting FE were annotated to the genus of *Streptococcus*. Fortunately, based on the existing database information, the most important candidate biomarker (OTU509) can also be annotated to species level, which was *Streptococcus gallolyticus* subsp. *gallolyticus*. Ten OTUs were annotated to the family of Lachnospiraceae (OTU962, OTU555, OTU1185, OTU931, OTU738, OTU684, OTU403, OTU399, OTU928, and OTU458). Two pairs of OTUs were annotated to the families of Erysipelotrichaceae (OTU1434 and OTU826) and Ruminococcaceae (OTU1094 and OTU1355). Four single OTUs were annotated to the families Coriobacteriaceae (OTU123), Peptococcaceae (OTU670), Prevotellaceae (OTU10), and Enterobacteriaceae (OTU398) (**Figure 3B** and **Supplementary Table S3**). The area under the ROC curve (AUG) was 0.99 based on the 24 most important variables (**Figure 3C**).

We further compared the abundance of OTUs between the high- and low-FE pigs using STAMP software with Welch's t-test. We detected only two OTUs (OTU509 and OTU1013) that were significantly different between pigs with low or high FE using p-value <0.05 and q-value <0.05 as the significance threshold. However, at a threshold of p-value <0.01 and q-value <0.1, we identified an additional six OTUs with a tendency toward a difference (**Figure 3D**). The average abundance of these OTUs between the high- and low-FE groups are shown in **Supplementary Figure S3**. Most of these OTUs were contained in the OTU list that outlined important variables to account for the differences in FE (**Supplementary Table S4**), except OTU1456, which was annotated to the order Clostridiales.

Comparison of the Functionality of the Fecal Microbiome in High- and Low-FE Groups Based on Metagenomic Sequencing

Comparison of the functional capacity of the gut microbiome can help to investigate the metabolic differences between high- and low-FE groups and further indicate the microbes that may affect special nutrient metabolism. The functional capacity was determined according to the annotation of ORFs predicted from the assembled contigs. The predicted genes were then aligned with the KEGG gene database to obtain the KO annotation

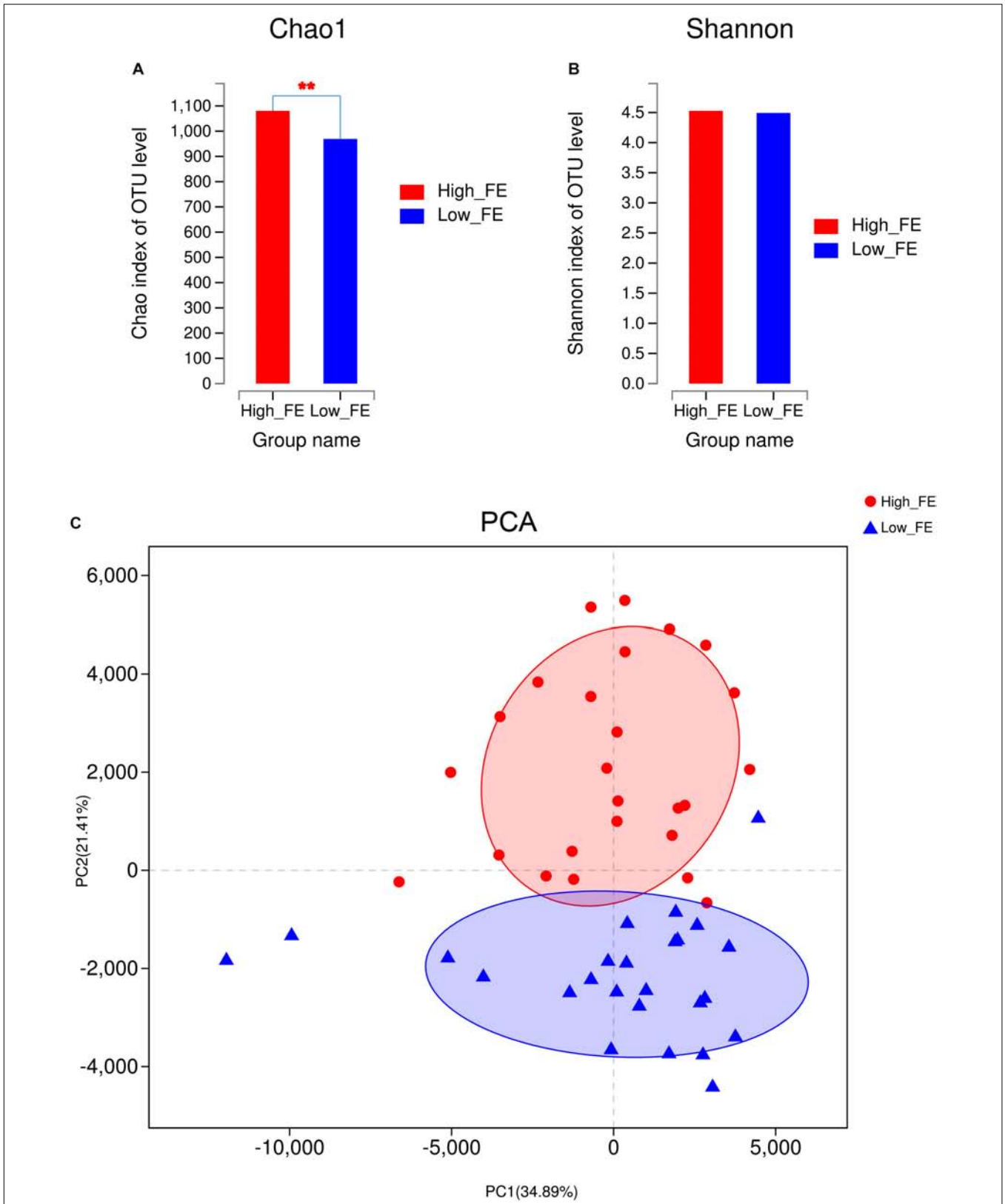


FIGURE 2 | The community diversity between fecal samples from high- and low-FE pigs at the OTU level. **(A)** The Chao1 index in high- and low-FE pigs. **(B)** The Shannon index in high- and low-FE pigs. **(C)** Principal component analysis (PCA) of the fecal microbiota based on OTUs.

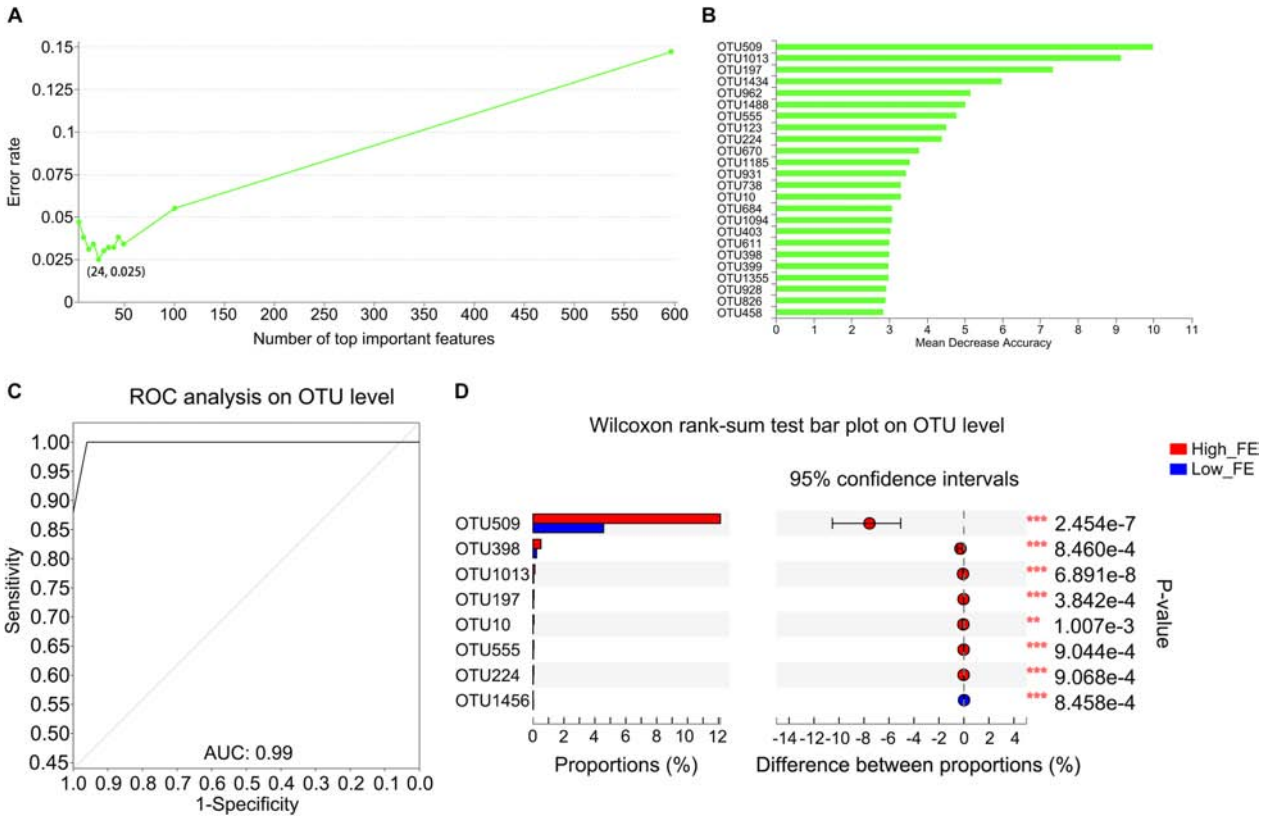


FIGURE 3 | Important biomarkers selected by random forest analysis and the different abundance of OTUs in high- and low-FE pigs. **(A)** Model evaluation by using top important variables. The horizontal axis represents the number of variables ranked by importance. The vertical axis denotes the average prediction error rate using 10-fold cross-validation when using the number of corresponding variables. **(B)** The ordination diagram of variables of importance. The horizontal axis is the measurement standard of variables of importance, and the value is equal to the measurement value of variables of importance/standard deviation. The vertical axis is the variable names sorted by importance. **(C)** ROC of the random forest classifier based on the 24 most important variables. The AUC value is the area under the corresponding curve. When the AUC > 0.5, the AUC value is closer to 1, the diagnostic effect is better. **(D)** The OTUs with a significantly different abundance between high- and low-FE pigs detected by STAMP software.

information from the KEGG database (see section "Materials and Methods"). A total of 1,857,107 ORFs were found with an average length of 616 bp. We identified a total of 352,002 KEGG genes and assigned them into 322 KEGG pathways. Subsequently, we compared the KEGG pathways abundance between the high- and low-FE groups, but no pathways were significantly different at FDR < 0.05. When we relaxed the threshold (p-value <0.05 and q-value <0.3), 18 pathways showed different enrichment at level 3. Eight pathways were more enriched in high-FE groups, and 10 pathways were more enriched in low-FE groups. The pathways that were enriched in high-FE pigs were associated with protein metabolism (ko04974), lipid metabolism (ko00600), and glycan degradation (ko00511). The pathways enriched in low-FE pigs involved endocrine regulation (ko03320 and ko04924), signal transduction (ko04152), the immune system (ko04622) and cardiovascular diseases (05410) (**Figure 4A** and **Supplementary Table S5**).

We further investigated the functional information of genes in the CAZy database; over forty thousand genes were identified and categorized into six CAZy classes. Glycoside hydrolases (GHs), glycosyl transferases (GTs) and carbohydrate-binding modules (CBMs) were the three classes enriched the most in both the high- and low-FE groups (**Supplementary Figure S4**). When we compared the EC activity abundance between the high- and low-FE groups, we found that 15 EC activities were more abundant in the high-FE groups, and 3 EC activities were enriched in the low-FE groups (p-value <0.05, q-value <0.05). The higher abundance of EC activities in the high-FE groups involved the degradation of xylan, cellulose and many other polysaccharides. The low-FE pigs have a greater abundance of three kinds of fucosyltransferases than the high-FE pigs (**Figure 4B** and **Supplementary Table S6**).

DISCUSSION

Metagenomic approaches based on high-throughput sequencing methods have rapidly facilitated the compositional and functional study of the gut microbiota in recent years (Fraher et al., 2012; Weinstock, 2012). Based on these high-throughput sequencing methods, many previous studies had revealed potential microbial biomarkers for improving FE in multiple

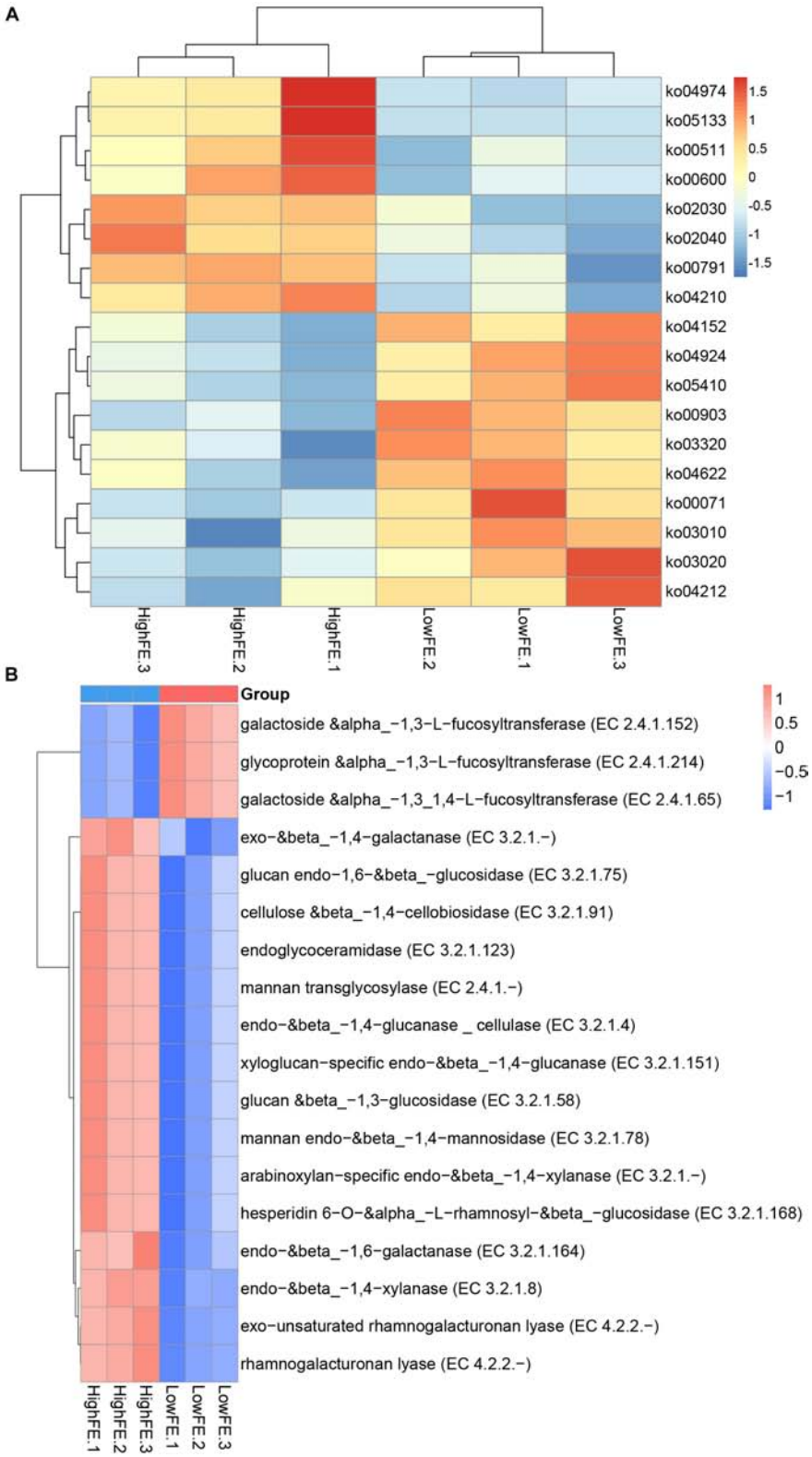


FIGURE 4 | Heatmap of functional capacity profiles showing different enrichment in high- and low-FE pigs by metagenomic sequencing analysis. Samples are coded according to the feed efficiency status (HighFE, high feed efficiency; LowFE, low feed efficiency). For example, HighFE.1 represented the first sample that was collected from high-feed-efficiency pig. **(A)** Heatmap of KEGG pathways showing different enrichments in high- and low-FE pigs. **(B)** Heatmap of CAZy EC activities showing different enrichments in high- and low.

breeds of pigs (McCormack et al., 2017; Tan et al., 2017, 2018; Yang et al., 2017; Quan et al., 2018). However, this study is one of the first to combine the technology of 16S rRNA gene sequencing and shotgun metagenomic sequencing to analyze the fecal microbial composition and function in commercial DLY pigs with high and low FE. All experimental pigs were selected from populations with similar genetic backgrounds. They were the same gender and were subjected to the same environmental, nutritional and management conditions to minimize the variability in FE due to genetic, gender, and external factors. Even in this well-controlled environment, there was still polarization in the FE of the experimental pig population. The difference in the intestinal microbiota in the experimental pigs was partly contributed to this phenomenon, which was suggested by previous studies (Nicholson et al., 2012; Parks et al., 2013; Yang et al., 2014). Although the experimental pig population and the number of sequencing samples is not particularly large in this study, the obvious FE variations reflect the real phenomenon in the pig industry. The pan-OTU number and the Good's coverage indices in the sequencing samples showed sufficient sampling of the population and adequate depth to investigate different bacterial species in the high- and low-FE pigs. The different bacterial species that were involved in feed nutrient processing and energy harvesting in high- and low-FE DLY pigs could be considered potential microbial biomarkers for FE.

When looking at the fecal microbiota composition, consistent with previous findings in pigs, the core phyla within the fecal microbiota were dominated by Firmicutes and Bacteroidetes (Yang et al., 2016; Xiao et al., 2017). Bacteroidetes have an important role in degrading indigestible dietary polysaccharides into short-chain fatty acids that can be reabsorbed by the host (Becker et al., 2014). Firmicutes were also thought to play a vital role in the energy harvest of mice (Turnbaugh et al., 2006). This finding may indicate that the dominant core phyla maintain a balance and can ensure the stability of intestinal function during the growth process in pigs. However, in our present study, we did not observe differences in *Bacteroides* and Firmicutes in high- and low-FE pigs. These findings differed with the results in pigs that showed an increase in Firmicutes in high fatness compared with low fatness subjects (Yang et al., 2016). Furthermore, when considering the annotation result at the genera level of the swine fecal microbiota, many studies had different classification compositions. In Yang et al.'s (2016) study, *Prevotella*, *Lactobacillus*, and *Treponema* were the three most abundant genera in Duroc pigs. *Prevotella*, *Streptococcus*, and *SMB53* were the three most abundant genera in Hampshire pigs, and *Clostridium*, *SMB53*, and *Streptococcus* were the three most abundant genera in Landrace and Yorkshire pigs in Xiao et al. (2017) study. This result suggested a special composition of the intestinal microbial community at the genus level, which may be due to differences in the breed, age, feed, and husbandry of pigs.

When we compared the bacterial community composition between the high- and low-FE pigs, we found that the community structure was significantly different (**Figures 2A,C**). The community of high-FE pigs had more richness and similar diversity to that of low-FE pigs. This finding suggested that the difference in FE is not due to the presence of specific bacteria in high-FE pigs but to the larger number of certain bacteria. These differences may come from colonized difference in the early life of mammalian, whose gut microbiome were thought to be at least partially shared by their parents, and it is relatively stable to perturbation once a dense microbial population is established (Antonopoulos et al., 2009; Snijders et al., 2016). However, no study has confirmed the causality between the microbial difference of offspring and their parents in pigs. In the present study, since the dams of our experimental pigs could not be fully tracked, we could not conclude that mother animals would cause a bias between high- and low FE pigs. In addition, gut microbiota composition may also be influenced by environmental factor (Yang et al., 2017), such as pig pen. However, in our study, the pig pen did not have a significant effect on porcine fecal microbial composition, but had a tendency to take a significant effect (p-value = 0.077) (**Supplementary Table S2**). In the current study, the random forest analysis showed that many OTUs played important roles in varying FE (**Figure 3B**). According to the annotation information from these OTUs, bacteria belonging to the genus *Streptococcus* and the families Lachnospiraceae, Erysipelotrichaceae, Ruminococcaceae, Coriobacteriaceae, Peptococcaceae, Prevotellaceae, and Enterobacteriaceae may be important candidates to improve swine FE. Furthermore, the OTUs that were enriched in high-FE pigs were mainly found in *Streptococcus*, *Escherichia-Shigella*, and Prevotellaceae *NK3B31* and the family Lachnospiraceae (**Figure 3D** and **Supplementary Table S4**).

A previous study suggested that Lachnospiraceae was associated with human obesity (Cho et al., 2012). Kameyama and Itoh (2014) reported that a bacterial strain of Lachnospiraceae can induce obesity in mice. Many members of Lachnospiraceae can produce short-chain fatty acids (SCFAs) by fermenting dietary polysaccharide (Pryde et al., 2002). The SCFAs were linked to a reduced risk of developing gastrointestinal disorders, cancer and cardiovascular disease and promote human obesity (Wong et al., 2006; Cho et al., 2012). Therefore, we hypothesized that the Lachnospiraceae might improve porcine FE by maintaining the gut in a healthy state to increase its absorptive capacity. Prevotellaceae is reported to relate to several diseases, such as asthmatic airway inflammation and arthritis, and associate with mucin degradation (Brinkman et al., 2011; Scher et al., 2013). Several members of Prevotellaceae were well-known succinate producer and can improve glucose homeostasis through activation of intestinal gluconeogenesis (De Vadder et al., 2016). A recent study reported that the succinate level was associated with carbohydrate metabolism and energy production (Serena et al., 2018). This study indicated that Prevotellaceae may increase FE in pigs by promoting host health or energy metabolism. A study in dairy calves suggested that SCFA concentration and carbohydrate utilization were significantly correlated with *Escherichia-Shigella* (Song et al., 2018). *Streptococcus* has been generally considered a health-promoting microbe for its roles in modulating human health (Kleerebezem and Vaughan, 2009). Many species belonging to *Streptococcus* were associated with carbohydrate fermentation, starch hydrolysis and the production of glucan

from sucrose (Facklam, 1972; Nelms et al., 1995). In this study, *Streptococcus gallolyticus* subsp. *gallolyticus* (annotated by OTU509) was considered an important candidate that might be used for improving porcine FE. Most strains of this species can ferment mannitol, trehalose, and inulin and can produce acid from starch and glycogen (Schlegel et al., 2003). These finding suggested that high-FE pigs are likely to have a greater abundance of intestinal microbes that can promote host intestinal health or degrade dietary carbohydrates. Therefore, high-FE pigs might have a greater ability to utilize feed and better intestinal health than low-FE pigs.

We also performed functional capacities analyses. Although we did not identify any core metabolic pathways at the q-value <0.05 level, some pathways showed a trend toward difference in high- and low-FE pigs. The pathways associated with protein metabolism (ko04974), lipid metabolism (ko00600), and glycan degradation (ko00511) were enriched in high-FE pigs. The higher abundance of protein metabolism and glycan degradation pathways in the high-FE groups had been reported in previous studies (Li and Guan, 2017; Yang et al., 2017). The experimental pigs in this study were fed with a fiber-enriched and high-protein diet. Therefore, the fecal microbiota may be more competent in terms of utilizing the diet protein. It is interesting that the fecal microbes of high-FE pigs have relatively more pathways of lipid metabolism, and it was believed that most of digestion and absorption occur in the small intestine (Rudd, 2012; Voet et al., 2013). We confirmed whether the fecal microbes have a compensatory metabolism function for unmetabolized lipid. Analysis of the microbial gene functional annotation in the CAZy database revealed expected results. In high-FE groups, the EC activities included the degradation of xylan, cellulose and many other polysaccharides. These functional results were consistent with the previous hypothesis that the high-FE pigs might have a greater ability to utilize dairy protein and carbohydrate than low-FE pigs.

CONCLUSION

In conclusion, there was a different microbial community structure in the fecal microbiota of pigs with different FE. We detected 24 OTUs that can serve as potential biomarkers to improve swine FE. Eight OTUs were significantly different or had a trend toward difference in the high- and low-FE pigs. The high-FE pigs had a greater abundance of OTUs in the families Lachnospiraceae and Prevotellaceae and in the genera *Escherichia-Shigella* and *Streptococcus* compared to low-FE pigs. *Streptococcus gallolyticus* subsp. *gallolyticus* could be an important candidate microbe for improving FE. We detected 18 KEGG pathways and CAZy EC activities that were different between high- and low-FE pigs. We found that the fecal microbiota in high-FE pigs have a greater capacity to degrade dietary cellulose, polysaccharide, and protein and may have a greater abundance of microbes to promote intestinal health. These findings should improve our understanding of the differences in the fecal microbial composition between high- and low-FE commercial pigs and provide important candidate microbes that can potentially use for improving porcine FE.

AUTHOR CONTRIBUTIONS

JY and ZW conceived and designed the experiments. JQ, MY, ZhoZ, RD, XW, ZhaZ, SZ, SL, HY, ZL, and EZ performed the experiments. JQ and JY analyzed the data. MY, GC, RD, XW, JY, EZ, and JQ collected the samples and recorded the phenotypes. GC and MY contributed the materials. JQ wrote the manuscript. JY, WH, and ZW revised the manuscript. All authors reviewed and approved the final manuscript.

FUNDING

This study was financially supported by the Natural Science Foundation of Guangdong Province (2018B030315007, 2018B030313011, and 2016A030310447) and Foundation of Department of Science and Technology of Guangdong Province of China (2015TX01N081), and the Foundation of Modern Agricultural Industrial Technology System of Guangdong Province (2018LM1101 and 2018LM1104).

ACKNOWLEDGMENTS

We would like to thank Guangdong Wen's Foodstuffs Group Co., for measuring the phenotypic traits and collecting the intestinal samples from the pigs. Thanks for the help provided by the free online platform, Majorbio I-Sanger Cloud Platform (www.i-sanger.com), in data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00052/full#supplementary-material

REFERENCES

Anderson, M., and Walsh, D. (2013). What null hypothesis are you testing? PERMANOVA, ANOSIM and the Mantel test in the face of heterogeneous dispersions. *Ecol. Monogr.* 83, 557–574.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* 26, 32–46.

Antonopoulos, D. A., Huse, S. M., Morrison, H. G., Schmidt, T. M., Sogin, M. L., and Young, V. B. (2009). Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. *Infect. Immun.* 77, 2367–2375. doi: 10.1128/IAI.01520-08

Becker, A. A., Hesta, M., Hollants, J., Janssens, G. P., and Huys, G. (2014). Phylogenetic analysis of faecal microbiota from captive cheetahs reveals underrepresentation of Bacteroidetes and Bifidobacteriaceae. *BMC Microbiol.* 14:43. doi: 10.1186/1471-2180-14-43

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

Brinkman, B. M., Hildebrand, F., Kubica, M., Goosens, D., Del Favero, J., Declercq, W., et al. (2011). Caspase deficiency alters the murine gut microbiome. *Cell Death Dis.* 2:e220. doi: 10.1038/cddis.2011.101

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

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, 335–336. doi: 10.1038/nmeth.f.303

Cho, I., Yamanishi, S., Cox, L., Methe, B. A., Zavadil, J., Li, K., et al. (2012). Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* 488, 621–626. doi: 10.1038/nature11400

De Vadder, F., Kovatcheva-Datchary, P., Zitoun, C., Duchampt, A., Backhed, F., and Mithieux, G. (2016). Microbiota-produced succinate improves glucose homeostasis via intestinal gluconeogenesis. *Cell Metab.* 24, 151–157. doi: 10.1016/j.cmet.2016.06.013

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Facklam, R. R. (1972). Recognition of group D streptococcal species of human origin by biochemical and physiological tests. *Appl. Microbiol.* 23, 1131–1139.

Ferrario, C., Alessandri, G., Mancabelli, L., Gering, E., Mangifesta, M., Milani, C., et al. (2017). Untangling the cecal microbiota of feral chickens by culturomic and metagenomic analyses. *Environ. Microbiol.* 19, 4771–4783. doi: 10.1111/1462-2920.13943

Fouhse, J. M., Gao, J., Vasanthan, T., Izydorczyk, M., Beattie, A. D., and Zijlstra, R. T. (2017). Whole-grain fiber composition influences site of nutrient digestion, standardized ileal digestibility of amino acids, and whole-body energy utilization in grower pigs. *J. Nutr.* 147, 29–36. doi: 10.3945/jn.116.238667

Fraher, M. H., O'Toole, P. W., and Quigley, E. M. (2012). Techniques used to characterize the gut microbiota: a guide for the clinician. *Nat. Rev. Gastroenterol. Hepatol.* 9, 312–322. doi: 10.1038/nrgastro.2012.44

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

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi: 10.1101/gr.112730.110

Hiergeist, A., Glasner, J., Reischl, U., and Gessner, A. (2015). Analyses of intestinal microbiota: culture versus sequencing. *ILAR J.* 56, 228–240. doi: 10.1093/ilar/ilv017

Kameyama, K., and Itoh, K. (2014). Intestinal colonization by a *Lachnospiraceae* bacterium contributes to the development of diabetes in obese mice. *Microbes Environ.* 29, 427–430. doi: 10.1264/jsme2.ME14054

Karlsson, F. H., Fak, F., Nookaew, I., Tremaroli, V., Fagerberg, B., Petranovic, D., et al. (2012). Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat. Commun.* 3:1245. doi: 10.1038/ncomms2266

Kleerebezem, M., and Vaughan, E. E. (2009). Probiotic and gut lactobacilli and bifidobacteria: molecular approaches to study diversity and activity. *Annu. Rev. Microbiol.* 63, 269–290. doi: 10.1146/annurev.micro.091208.073341

Li, F., and Guan, L. L. (2017). Metatranscriptomic profiling reveals linkages between the active rumen microbiome and feed efficiency in beef cattle. *Appl. Environ. Microbiol.* 83:e00061-17. doi: 10.1128/AEM.00061-17

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics* 24, 713–714. doi: 10.1093/bioinformatics/btn025

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18. doi: 10.1186/2047-217X-1-18

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

McCormack, U. M., Curiao, T., Buzoianu, S. G., Prieto, M. L., Ryan, T., Varley, P., et al. (2017). Exploring a possible link between the intestinal microbiota and feed efficiency in pigs. *Appl. Environ. Microbiol.* 83:e00380-17. doi: 10.1128/AEM.00380-17

Nelms, L. F., Odelson, D. A., Whitehead, T. R., and Hespell, R. B. (1995). Differentiation of ruminal and human *Streptococcus bovis* strains by DNA homology and 16s rRNA probes. *Curr. Microbiol.* 31, 294–300.

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., et al. (2012). Host-gut microbiota metabolic interactions. *Science* 336, 1262–1267. doi: 10.1126/science.1223813

Parks, B. W., Nam, E., Org, E., Kostem, E., Norheim, F., Hui, S. T., et al. (2013). Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metab.* 17, 141–152. doi: 10.1016/j.cmet.2012.12.007

Pryde, S. E., Duncan, S. H., Hold, G. L., Stewart, C. S., and Flint, H. J. (2002). The microbiology of butyrate formation in the human colon. *FEMS Microbiol. Lett.* 217, 133–139.

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. doi: 10.1038/nature11450

Quan, J., Cai, G., Ye, J., Yang, M., Ding, R., Wang, X., et al. (2018). A global comparison of the microbiome compositions of three gut locations in commercial pigs with extreme feed conversion ratios. *Sci. Rep.* 8:4536. doi: 10.1038/s41598-018-22692-0

Rotz, C. A. (2004). Management to reduce nitrogen losses in animal production. *J. Anim. Sci.* 82(E-Suppl.), E119–E137.

Rudd, D. (2012). Elsevier's integrated review biochemistry 2nd edition. *Aust. J. Med. Sci.* 33:182.

Scher, J. U., Sczesnak, A., Longman, R. S., Segata, N., Ubeda, C., Bielski, C., et al. (2013). Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* 2:e01202. doi: 10.7554/eLife.01202

Schlegel, L., Grimont, F., Ageron, E., Grimont, P. A., and Bouvet, A. (2003). Reappraisal of the taxonomy of the *Streptococcus bovis*/*Streptococcus equinus* complex and related species: description of *Streptococcus gallolyticus* subsp. *gallolyticus* subsp. nov., *S. gallolyticus* subsp. *macedonicus* subsp. nov. and *S. gallolyticus* subsp. *pasteurianus* subsp. nov. *Int. J. Syst. Evol. Microbiol.* 53(Pt 3), 631–645. doi: 10.1099/ijs.0.02361-0

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 6:e27310. doi: 10.1371/journal.pone.0027310

Serena, C., Ceperuelo-Mallafre, V., Keiran, N., Queipo-Ortuno, M. I., Bernal, R., Gomez-Huelgas, R., et al. (2018). Elevated circulating levels of succinate in human obesity are linked to specific gut microbiota. *ISME J.* 12, 1642–1657. doi: 10.1038/s41396-018-0068-2

Snijders, A. M., Langley, S. A., Kim, Y. M., Brislawn, C. J., Noecker, C., Zink, E. M., et al. (2016). Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome. *Nat. Microbiol.* 2:16221. doi: 10.1038/nmicrobiol.2016.221

Song, Y., Malmuthuge, N., Steele, M. A., and Guan, L. L. (2018). Shift of hindgut microbiota and microbial short chain fatty acids profiles in dairy calves from birth to pre-weaning. *FEMS Microbiol. Ecol.* 94:fix179. doi: 10.1093/femsec/fix179

Tan, Z., Wang, Y., Yang, T., Ao, H., Chen, S., Xing, K., et al. (2018). Differences in gut microbiota composition in finishing Landrace pigs with low and high feed conversion ratios. *Antonie Van Leeuwenhoek* 111, 1673–1685. doi: 10.1007/s10482-018-1057-1

Tan, Z., Yang, T., Wang, Y., Xing, K., Zhang, F., Zhao, X., et al. (2017). Metagenomic analysis of cecal microbiome identified microbiota and functional capacities associated with feed efficiency in landrace finishing pigs. *Front. Microbiol.* 8:1546. doi: 10.3389/fmicb.2017.01546

Teagasc (2015). *Pig Herd Performance Report 2015*. Foster City, CA: Teagasc Pig Development Department.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

Voet, D., Voet, J. G., and Pratt, C. W. (2013). *Fundamentals of Biochemistry:Life at the Molecular Level*, Fourth Edn. Hoboken, NJ: Wiley.

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

Weinstock, G. M. (2012). Genomic approaches to studying the human microbiota. *Nature* 489, 250–256. doi: 10.1038/nature11553

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

Wong, J. M., de Souza, R., Kendall, C. W., Emam, A., and Jenkins, D. J. (2006). Colonic health: fermentation and short chain fatty acids. *J. Clin. Gastroenterol.* 40, 235–243.

Xiao, L., Estelle, J., Kiilerich, P., Ramayo-Caldas, Y., Xia, Z., Feng, Q., et al. (2016). A reference gene catalogue of the pig gut microbiome. *Nat. Microbiol.* doi: 10.1038/nmicrobiol.2016.161 [Epub ahead of print].

Xiao, Y., Li, K., Xiang, Y., Zhou, W., Gui, G., and Yang, H. (2017). The fecal microbiota composition of boar Duroc, Yorkshire, Landrace and Hampshire pigs. *Asian Australas. J. Anim. Sci.* 30, 1456–1463. doi: 10.5713/ajas.16.0746

Yang, H., Huang, X., Fang, S., He, M., Zhao, Y., Wu, Z., et al. (2017). Unraveling the fecal microbiota and metagenomic functional capacity associated with feed efficiency in pigs. *Front. Microbiol.* 8:1555. doi: 10.3389/fmicb.2017.01555

Yang, H., Huang, X., Fang, S., Xin, W., Huang, L., and Chen, C. (2016). Uncovering the composition of microbial community structure and metagenomics among three gut locations in pigs with distinct fatness. *Sci. Rep.* 6:27427. doi: 10.1038/srep27427

Yang, L., Bian, G., Su, Y., and Zhu, W. (2014). Comparison of faecal microbial community of lantang, bama, erhualian, meishan, xiaomeishan, duroc, landrace, and yorkshire sows. *Asian Australas. J. Anim. Sci.* 27, 898–906. doi: 10.5713/ajas.2013.13621

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 Statement: 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.

Copyright © 2019 Quan, Cai, Yang, Zeng, Ding, Wang, Zhuang, Zhou, Li, Yang, Li, Zheng, Huang, Yang and Wu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 31 January 2019
doi: 10.3389/fmicb.2018.03333

Semen Microbiome Biogeography: An Analysis Based on a Chinese Population Study

Zhanshan Ma[1,2]* *and Lianwei Li*[1]

[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

Investigating inter-subject heterogeneity (or spatial distribution) of human semen microbiome diversity is of important significance. Theoretically, the spatial distribution of biodiversity constitutes the core of *microbiome biogeography*. Practically, the inter-subject heterogeneity is crucial for understanding the normal (healthy) flora of semen microbiotas as well as their possible changes associated with abnormal fertility. In this article, we analyze the scaling (changes) of semen microbiome diversity across individuals with DAR (diversity-area relationship) analysis, a recent extension to classic SAR (species-area relationship) law in biogeography and ecology. Specifically, the unit of "area" is individual subject, and the microbial diversity in seminal fluid of an individual (area) is assessed *via* metagenomic DNA sequencing technique and measured in the Hill numbers. The DAR models were then fitted to the accrued diversity across different number of individuals (area size). We further tested the difference in DAR parameters among the healthy, subnormal, and abnormal microbiome samples in terms of their fertility status based on a cross-sectional study of a Chinese cohort. Given that no statistically significant differences in the DAR parameters were detected among the three groups, we built unified DAR models for semen microbiome by combining the healthy, subnormal, and abnormal groups. The model parameters were used to (i) estimate the microbiome diversity scaling in a population (cohort), and construct the so-termed DAR profile; (ii) predict/construct the maximal accrual diversity (MAD) profile in a population; (iii) estimate the pair-wise diversity overlap (PDO) between two individuals and construct the PDO profile; (iv) assess the ratio of individual diversity to population (RIP) accrual diversity. The last item (RIP) is a new concept we propose in this study, which is essentially a ratio of local diversity to regional or global diversity (LRD/LGD), applicable to general biodiversity investigation beyond human microbiome.

Keywords: semen microbiome, biogeography, inter-subject heterogeneity, DAR (diversity-area relationship), beta-diversity, RIP (ratio of individual to population accrual diversity), LRD/LGD (ratio of local to regional/global diversity)

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Lei Hou,
Massachusetts Institute of Technology, United States
Dongya Jia,
Rice University, 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: *08 October 2018*
Accepted: *24 December 2018*
Published: *31 January 2019*

Citation:
Ma Z and Li L (2019) Semen Microbiome Biogeography: An Analysis Based on a Chinese Population Study. Front. Microbiol. 9:3333. doi: 10.3389/fmicb.2018.03333

INTRODUCTION

Similar to other human microbiome habitats such as gut, vaginal, or breast milk, human seminal fluid also hosts a microbiome including several hundreds of bacterial species per individual with various levels of abundances (Kiessling et al., 2008; Moretti et al., 2009; De Francesco et al., 2011; Hou et al., 2013; Weng et al., 2014). The seminal microbiome, just like other human microbiomes, is highly personalized. If we considered a cohort or population of men, their semen microbiomes are independent in ecological time in general, and each individual is not unlike an island available for microbes to invade and/or inhabit. Similar scenarios have been investigated extensively in macro-ecology of plants and animals since the 1960s, started with MacArthur and Wilson's (1967) island biogeography. The biogeography studies the spatial and/or temporal distribution of biodiversity and is a foundation of the modern conservation biology and biodiversity conservation in large. It has been widely recognized that seminal microbiome is implicated, at least in some of the male infertilities (Kiessling et al., 2008; Moretti et al., 2009; De Francesco et al., 2011; Domes et al., 2012; Hou et al., 2013; Weng et al., 2014). Therefore, investigating the biogeography or spatial distribution of seminal microbiome diversity is necessary for deep understanding the seminal microbiome as well as their implications for male infertility.

Prior to recent large-scale DNA sequencing studies of seminal microbiome samples (e.g., Hou et al., 2013; Weng et al., 2014), most studies on seminal microbes were focused on acute and chronic microbial infections, either based on PCR, microscopic or artificially culture-based methods (Keck et al., 1998; Henkel et al., 2006; Kiessling et al., 2008; Lbadin and Ibeh, 2008; Ochsendorf, 2008; Moretti et al., 2009; Akutsu et al., 2012; Domes et al., 2012), and majority of the early studies were conducted to explore the relationship between infections and male infertility. It was reported that infectious etiologies cause about 15% of male infertility cases (Diemer et al., 2003; Weng et al., 2014). The adoption of NGS (next generation sequencing) technologies have lead to significant advances in understanding the semen microbiome, because it greatly expanded our capability to detect virtually all bacteria in seminal fluid with rather low cost. Since the cataloging the semen microbes is not limited to infectious or opportunistically infectious microbes anymore, the NGS-based metagenomic technology and associated bioinformatics analyses have made the examination of the whole seminal microbiome from ecological perspective a routine research technique. For example, Weng et al. (2014) showed that the most abundant genera among the semen samples of 96 Chinese individuals were *Lactobacillus* (19.9%), *Pseudomonas* (9.85%), *Prevotella* (8.51%), and *Gardnerella* (4.21%). They further found that the seminal bacterial communities were clustered (through unsupervised clustering analysis) into three major types, dominated by *Lactobacillus*, *Pseudomonas*, and *Prevotella*, respectively. They also investigated the association between seminal microbial community and semen quality. In spite of the significant advances made in the existing studies, to the best of our knowledge, no studies with biogeography approaches to seminal fluid microbiome have ever been performed. As mentioned previously, biogeography approaches offer applicable theory and ideal techniques for analyzing the spatial distribution patterns of seminal microbiome diversity in a human population or cohort, and insights from the biogeography approaches such as heterogeneities of the seminal microbiome among individuals and the population-level characteristics should certainly be rather useful for personalized fertility research and public health.

Microbial biogeography is charged with the mission of understanding the spatial and/or temporal distribution of microbial diversities on regional or global scales. The classic species-area relationship (SAR), which quantitatively characterizes the relationship between the number of species (formally known as *species richness*, which is a rough measure of biodiversity) and the geographic area species distributed as a power-law function, is regarded as one of few classic laws in ecology and biogeography. The first documentation of the SAR relationship can be traced back to British botanist (Watson's, 1835) study of the distribution of plants. Since then, numerous theoretical and field studies have been performed (Watson, 1835; Preston, 1960; Connor and McCoy, 1979; Rosenzweig, 1995; Harte et al., 1999; Lomolino, 2000; Drakare et al., 2006; Tjørve and Tjørve, 2008; Tjørve, 2009; He and Hubbell, 2011; Sizling et al., 2011; Storch et al., 2012; Triantis et al., 2012; Whittaker and Triantis, 2012; Helmus et al., 2014). In the 1960's, the SAR theory inspired (MacArthur and Wilson's, 1967) establishment of their island biogeography theory, and the theory not only greatly enriched the principles and methods of general biogeography, but also was essential in shifting the focus of ecological research from population to community and in advancing *community ecology* in the 1970s and after. Today, much of the ecological theories and analysis techniques applied to microbiome research come from *community ecology*.

Recently, taking advantage of the big metagenomic datasets from the human microbiome project (HMP) and related studies, Ma, 2018a,b extended the classic SAR to general DAR (diversity-area relationship) by replacing the "species richness" in the classic SAR with general "diversity." As mentioned previously, *species richness* or the number of species in a community, region or area, is rather rough as a measure of biodiversity because it ignores the fact that not all species are born equally abundant on the planet. Some species like panda are on one extreme and others such as flies are another extreme. The classic SAR is therefore somewhat flawed when the relationship is used to characterize the spatial distribution of biodiversity thanks to the simplified measure of biodiversity with species numbers. The DAR overcomes the flaw of traditional SAR by using more scientific metrics for biodiversity measures. Specifically, to construct DAR models, Ma, 2018a,b utilized Renyi's entropy based Hill numbers, which included some of the most widely used diversity indexes such as Shannon diversity and Simpson diversity indexes as its special cases. The adoption of Hill numbers for building DAR models also overcomes an issue of selecting diversity index from many of the choices, which often confuses non-ecologists unnecessarily.

The present article aims to apply the recent extended DAR modeling approach to discovering the important patterns of biogeography of seminal microbiome. We build the DAR models and compute these metrics by using the metagenomic

sequencing data originally reported by Weng et al. (2014), and we also explore whether or not those metrics are related to the sperm quality. Specifically, we build DAR models for alpha-diversity and beta-diversity, respectively, and further derive some critical parameters including *diversity scaling parameter*—measuring the change rates of diversity across individuals (the size of microbial habitat area), *pair-wise diversity overlap* (similarity) (PDO)—measuring the average proportion of shared diversity or similarity between two individuals, *maximum accrual diversity* (MAD) in a population or cohort, and the ratio of individual to population diversity (*RIP*), a newly introduced metric that measures the ratio of individual microbial diversity to population-level microbial diversity. In terms of more general biogeography terms beyond the human microbiome, the concept of RIP can be generalized as the ratio of local to regional diversity (LRD) or ratio of local to global diversity (LGD), which can be applied to general biodiversity research in any other ecosystems.

MATERIALS AND METHODS

Datasets Description

The 16S-rRNA OTU (operational taxonomic unit) tables of the semen microbiome at genus and species taxonomic levels, respectively, which we used to perform the DAR analysis, were originally reported by Weng et al. (2014). The OTU tables were generated from DNA-sequencing the semen microbiome samples, collected from 96 individuals including 35 with normal fertility, 28 with sub-normal fertility, and 33 with abnormal fertility, and the consequent bioinformatics analysis. From the 96 samples, Weng et al. (2014) obtained a total of 8,337,766 sequence reads, that is 80,424 reads per participant sample, a sufficiently large sample size for consequent statistical analyses. They detected an average number of 135 genera and 569 species from those samples.

Since the objective of Weng et al. (2014) study was to investigate the relationship between sperm quality and seminal microbiome, the original study included three treatments (groups), i.e., the normal, sub-normal, and abnormal as mentioned previously. The study design, of course, has no issue at all with its original objectives. To harness the data for our DAR analysis in this study, we first build DAR models for each treatment separately, and then perform statistical tests to see if there are any differences in the DAR parameters among the three treatments. If there is any significant difference, we keep the results and further investigate the implications of the difference to the status of treatments (fertility status). If there is not any significant difference, we then combine all 96 samples from the three treatments, build a single set of DAR models with the combined datasets, and further use the DAR models to explore the general biogeography properties of the seminal microbiome.

The Diversity-Area Relationship (DAR)

The process of constructing DAR models for microbes consists of the following three steps: (i) bioinformatics analysis of 16S-rRNA reads to get OUT tables (Schloss et al., 2009; Caporaso et al., 2010; Bokulich et al., 2018); (ii) computing species or OTU diversities measured with the Hill numbers (Chao et al., 2012, 2014; Ma, 2017); (iii) building the DAR models (Ma, 2018a,b).

Diversity Measured in Hill Numbers

The Hill numbers are a form of Renyi's entropy (Renyi, 1961). It was initially introduced as an *evenness* index from economics by Hill (1973) and later reintroduced into ecology by Jost (2007) and Chao et al. (2012) who further clarified Hill's numbers for measuring alpha diversity as:

$$ {}^{q}D = \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)} \tag{1} $$

where S is the number of species, p_i is the relative abundance of species i, q is the order number of diversity.

The Hill number is undefined when $q = 1$, but its limit as q approaches to *1* exists in the following form:

$$ {}^{1}D = \lim_{q \to 1} {}^{q}D = \exp\left(- \sum_{i=1}^{s} p_i \log(p_1) \right) \tag{2} $$

The parameter q controls the sensitivity of the Hill number to the relative frequencies of species abundances. When $q = 0$, the species abundances do not weigh at all and ${}^{0}D = S$, i.e., species richness. When $q = 1$, ${}^{1}D$ equal the *exponential* of Shannon entropy, and is interpreted as the number of typical or common species in the community because ${}^{1}D$ is weighted proportionally by species abundances. When $q = 2$, ${}^{2}D$ equal the reciprocal of Simpson index, i.e.,

$$ {}^{2}D = (1/\sum_{i=1}^{S} p_i^2) \tag{3} $$

which is interpreted as the number of *dominant* or very abundant species in the community (Chao et al., 2012) because ${}^{2}D$ is weighted in favor of more abundant species. The general interpretation of ${}^{q}D$ is that the community has a diversity of order q, which is equivalent to the diversity of a community with ${}^{q}D = x$ equally abundant species.

A recent consensus suggested that, with the Hill numbers, the multiplicatively defined beta-diversity, rather than additively defined, by partitioning gamma diversity into the product of alpha and beta, should be used to define beta-diversity, in which both alpha and gamma diversities are measured with the Hill numbers.

$$ {}^{q}D_{\beta} = {}^{q}D_{\gamma} / {}^{q}D_{\alpha} \tag{4} $$

This beta diversity derived from the above partition takes the value of *1* if all communities are identical, the value of N (the number of communities) when all the communities are completely different from each other (there are no shared species). With Jost (2007) words, this beta diversity measures "*the effective number of completely distinct communities*." In this study, we compute diversities until $q = 3$, i.e., to the third order. Note that a series of the Hill numbers at different order q is termed *diversity profile* (Jost, 2007; Chao et al., 2012, 2014).

The DAR (Diversity-Area Relationship) Models

Based on the fact that all Hill numbers are in the units of species or species equivalents such as OTUs, and on the intuition that Hill numbers should follow the same or similar pattern of the classic SAR (species area relationship), Ma (2018a) extended SAR to general DAR (diversity-area relationship), in which diversity is measured with Hill numbers.

The basic power function, known as the *power law* (PL) species scaling law widely adopted in SAR study, is extended to describe the general diversity-area relationship (DAR):

$$^qD = cA^z \tag{5}$$

where ^qD is diversity measured in the *q-th* order Hill numbers, A is *area*, and c and z are parameters.

A slightly modified PL model, the power law with exponential cutoff (PLEC) model, originally introduced to SAR modeling by Plotkin et al. (2000) and Ulrich and Buszko (2003), respectively (also see Tjørve, 2009), can also be utilized for DAR modeling. The PLEC model is:

$$^qD = cA^z \exp(dA), \tag{6}$$

where d is a third parameter and should be negative in DAR scaling models, and $\exp(dA)$ is the exponential decay term that eventually overwhelms the power law behavior at very large value of A.

The following log-linear transformed equations (7, 8) can be used to estimate the model parameters of Equations (5, 6), respectively:

$$\ln(D) = \ln(c) + z\ln(A) \tag{7}$$

$$\ln(D) = \ln(c) + z\ln(A) + dA \tag{8}$$

Both linear correlation coefficient (R) and p-value can be used to judge the goodness of the model fitting. In fact, either of them should be sufficient to judge the suitability of the models to data. Three advantages are associated with the linear-transformed fitting: (i) simplicity in computation, (ii) parameter z is scale-invariant with Equation (7), (iii) the ecological interpretation of scaling parameter is preserved with Equation (8). The scaling parameter z is also termed the *slope* of the DAR power-law, because z represents the *slope* of the linearized function in log–log space.

The relationship between DAR model parameter (z) of the DAR PL model and the diversity order (q), or z-q trend, was defined as the *DAR profile* (Ma, 2018a). It describes the change of diversity scaling parameter (z) with the diversity order (q), comprehensively. Obviously, the DAR profile is an extension of the *diversity profile* concept Chao et al. (2012, 2014) proposed, which is the diversity in the Hill numbers at the q-th order.

In macro-ecology, there are usually natural spatial orders for the "areas," which is generally lacking in human microbiome because human residences are little relevant to the accrual of diversities for DAR modeling. To avoid the potential bias from an arbitrary order of the human microbiome samples, we totally permutated the orders of all the microbiome samples under investigation, and then randomly chose 100 orders of the microbiome samples generated from the total permutations. That is, rather than taking a single arbitrary order for accruing microbiome samples in one-time fitting to the DAR model, we repeatedly perform the DAR model-fitting 100 times with the 100 randomly chosen permutation orders. Finally, the averages of the model parameters from the 100 times of DAR fittings are adopted as the model parameters of the DAR for the set of microbiome samples under investigation. An additional advantage of this re-sampling from total permutations is that the procedure makes the parameter c of the DAR-PL model being able to represent an average individual in the population (cohort) from which the individual comes from.

Predicting MAD (Maximal Accrual Diversity) With DAR-PLEC Models

Ma (2018a) derived the maximal accrual diversity (MAD) in a cohort (or population) based on the PLEC model [Equations (6, 8)] as follows:

$$Max(^qD) = {}^qD_{\max} = c(-\frac{z}{d})^z \exp(-z) = cA^z_{\max}\exp(-z) \tag{9}$$

and the number of individuals ($A_{\max}$) needed to reach the maximum can be estimated by

$$A_{\max} = -z/d \tag{10}$$

where all parameters are the same as Equations (6,8).

Similar to the previous definition for DAR profile (z-q pattern), the MAD profile ($D_{\max}$-q pattern), was defined as a series of $D_{\max}$ values corresponding to different diversity order (q) (Ma, 2018a).

Pair-Wise Diversity Overlap (PDO) Profile

The *pair-wise diversity overlap* (g) of two bordering areas of the same size (i.e., the proportion of the new diversity in the second area) is (Ma, 2018a):

$$g = 2 - 2^z \tag{11}$$

where z is the scaling parameter of DAR-PL model [Equations (5, 7)]. If $z = 1$, then $g = 0$, there is no overlap (similarity); if $z = 0$, then $g = 1$, totally overlap. In reality, g should be between 0 and 1.

Since the *equal size* of *area* assumption is largely true in the case of sampling human microbiome, the parameter z of the DAR-PL can be utilized to estimate the *pair-wise diversity overlap* (PDO), i.e., the diversity overlap (similarity) between two individuals, in the human microbiome with Equation (11).

Similar to previous definitions for DAR profile (z-q pattern) and MAD profile ($D_{\max}$-q pattern), the PDO profile (g-q pattern) was defined as a series of PDO-g values at different diversity order (q) (Ma, 2018a).

A Summary on the Interpretations of Important DAR Parameters

We summarize the ecological interpretations from PL/PLEC as follows to facilitate the discussion of the results from fitting DAR models with semen microbiome datasets.

z: The slope of the DAR-PL model or *scaling* parameter, and it is the ratio of diversity accrual rate to area increase rate. The *DAR profile* is a series of z-q values, corresponding to different diversity order (q).

c: Theoretically, by setting $A = 1$, $S_0 = cA^z = c$, hence c is the Hill numbers (i.e., the number of species or species equivalents of diversity) in *one unit of area,* but not *per unit of area* because the scaling is non-linear. However, since we used 100 times of re-sampling to get the DAR parameters as explained previously, plus that the area size in human microbiome sampling can be considered as approximately equal, we argue that, in practice, the parameter c of the DAR-PL model can be treated as an estimate of the individual-level diversity in Hill numbers, or of the diversity of an averaged individual in the cohort (or population) he or she belongs to.

g: The pair-wise diversity overlap (PDO) parameter. It measures the pair-wise diversity similarity between two neighboring areas of the same size, i.e., between two individuals in a cohort (or population). The *PDO profile* is a series of g-q values, corresponding to different diversity order (q).

D_{max}: The maximal accrual diversity (MAD) parameter. It estimates the maximal accrual diversity across individuals. Theoretically, it should be specific to the microbiome type (e.g., the gut microbiome or semen microbiome). The *MAD profile* is a series of D_{max}-q values, corresponding to different diversity order (q).

RIP (the Ratio of Individual Diversity to Population Accrual Diversity)—A New Definition

We define the RIP (Ratio of Individual diversity to Population accrual diversity) as:

$$^qRIP = {}^qc/{}^qD \tag{12}$$

where ${}^qc$ is the DAR-PL parameter at diversity order of q, and ${}^qD$ is the estimated accrual diversity of the population (cohort) with DAR-PL model at diversity order of q.

We further define ${}^qRIP$-q series (there is a *RIP* for each diversity order q) as RIP profile, similar to the previously defined DAR-, PDO-, and MAD-profiles.

According to the above RIP definition, a RIP profile can be constructed with population (cohort) of any size. However, in practice, using ${}^qD_{max}$ in place of ${}^qD$ should be more convenient, that is:

$$^qRIP = {}^qc/{}^qD_{\max} \tag{13}$$

The RIP parameter measures the average level of an individual can represent a population (or cohort) from which the individual comes from. As argued previously, parameter c is an *approximated* value of individual diversity (or diversity *per* individual). The approximation is contingent on two implicit assumptions: (i) the sizes of areas are equal, which is generally true in the case of human microbiome; (ii) the start of area accrual won't exert significant influence on the estimation of parameter c. This appears to be satisfied given that assumption (i) is largely true for the human microbiome. However, given the well-known inter-individual heterogeneity of the human microbiome, the choice of starting area (individual) to accrue diversity may indeed have a significant impact on the estimate of parameter c. To deal with the issue associated with assumption (ii), we adopt the previously introduced the re-sampling approach from total permutations of the microbiome samples, and use the average parameters from certain times (usually 100 should be enough) of repeatedly DAR model-fitting from the re-sampling.

In general biogeography terms beyond human microbiome, the previous definitions for RIP can be generalized as LRD (ratio of local to regional diversity) (Equation 14) or as LGD (ratio of local to global diversity (Equation 15). Both can be applied to measure the relationship between the local and regional/global biodiversities in any ecosystems. LRD & LGD are defined as:

$$^qLRD = {}^qc/{}^qD \tag{14}$$

$$^qLGD = {}^qc/{}^qD_{\max} \tag{15}$$

where the symbols (parameters) in the right have the same interpretations as in Equations (13,14).

RESULTS AND DISCUSSION

Test the Differences in Semen DAR Parameters Among the Three Groups

We aimed to test whether or not there are significant differences among the three groups (normal, sub-normal and abnormal) in their DAR parameters. To perform this test, we built DAR models (including both alpha-DAR and beta-DAR models) for each group separately and then performed the randomization tests for the parameters of those DAR models. The parameters of the alpha-DAR models and beta-DAR models for the three different groups were listed in **Tables S1, S2** of the online supplementary information (OSI), respectively. The results from the randomization test for the model parameters were listed in **Table S3** (for alpha-DAR parameters) and **Table S4** (for beta-DAR parameters), respectively. It turned out that there were no significant differences in any of the major DAR parameters between the groups, as revealed by the p-values ($p > 0.05$) in the last column of **Tables S3, S4**.

Biogeography Analysis of the Semen Microbiome With DAR Modeling

Alpha-DAR Modeling

Tables 1, **2** listed the alpha-DAR parameters for the human semen microbiome at the genus and species level, respectively. The leftmost column in both the tables listed the diversity order ($q = 0, 1, 2, 3$) and, and the parameters for DAR-PL models and DAR-PLEC models were listed in the left and right side, respectively. From **Tables 1**, **2**, we summarize the following findings:

(i) The DAR models fitted to the semen microbiome diversity in the Hill numbers at both genus and species levels statistically significant ($p < 0.05$ in 6 cases and $p < 0.1$ in two cases). Judged from the success rates among 100 times

TABLE 1 | The parameters of *alpha*-DAR (*alpha*-diversity-area relationship) computed with 100 times of re-sampling at **genus** level for the human semen microbiome.

Diversity order and statistics		Power law (PL)						PL with exponential cutoff (PLEC)							
		z	ln(*c*)	*R*	*p*-value	*g*	*N**	*z*	*d*	ln(*c*)	*R*	*p*-value	*N*	A_{max}	D_{max}
$q = 0$	Mean	0.277	5.148	0.984	0.000	0.788	100	0.344	−0.002	5.027	0.993	0.000	100	139	590.8
	Std. Err.	0.031	0.129	0.012	0.000	0.026		0.065	0.002	0.181	0.006	0.000			
	Min	0.211	4.812	0.936	0.000	0.718		0.218	−0.006	4.533	0.963	0.000			
	Max	0.359	5.444	0.998	0.000	0.843		0.514	0.000	5.384	0.999	0.000			
$q = 1$	Mean	0.100	3.176	0.668	0.039	0.927	94	0.154	−0.002	3.078	0.761	0.022	95	77	36.3
	Std. Err.	0.076	0.318	0.250	0.168	0.056		0.131	0.003	0.377	0.215	0.112			
	Min	−0.049	2.303	0.008	0.000	0.774		−0.230	−0.010	2.005	0.061	0.000			
	Max	0.294	3.832	0.975	0.940	1.034		0.546	0.010	4.051	0.987	0.839			
$q = 2$	Mean	0.075	2.308	0.514	0.061	0.944	89	0.125	−0.002	2.217	0.667	0.013	97	67	13.7
	Std. Err.	0.107	0.463	0.250	0.208	0.080		0.170	0.005	0.490	0.200	0.084			
	Min	−0.133	1.013	0.003	0.000	0.721		−0.365	−0.011	0.723	0.069	0.000			
	Max	0.355	3.171	0.937	0.976	1.088		0.639	0.017	3.342	0.968	0.799			
$q = 3$	Mean	0.055	1.998	0.455	0.093	0.958	76	0.102	−0.002	1.914	0.629	0.030	92	59	9.3
	Std. Err.	0.111	0.483	0.268	0.208	0.082		0.172	0.005	0.489	0.216	0.123			
	Min	−0.161	0.719	0.000	0.000	0.744		−0.385	−0.012	0.439	0.073	0.000			
	Max	0.329	2.877	0.942	0.999	1.105		0.602	0.017	2.943	0.959	0.780			

N, the number of successful fitting to DAR model from 100 times of random re-sampling of the individual orders.*

TABLE 2 | The parameters of *alpha*-DAR (*alpha*-diversity-area relationship) computed with 100 times of re-sampling at **species** level for the human semen microbiome.

Diversity order and statistics		Power law (PL)						PL with exponential cutoff (PLEC)							
		z	ln(*c*)	*R*	*p*-value	*g*	*N**	*z*	*d*	ln(*c*)	*R*	*p*-value	*N*	A_{max}	D_{max}
$q = 0$	Mean	0.507	6.627	0.984	0.000	0.577	100	0.637	−0.005	6.393	0.993	0.000	100	132	7099.5
	Std. Err.	0.061	0.247	0.012	0.000	0.060		0.127	0.003	0.354	0.006	0.000			
	Min	0.341	5.937	0.937	0.000	0.400		0.334	−0.012	5.357	0.962	0.000			
	Max	0.678	7.347	0.998	0.000	0.734		1.000	0.002	7.213	0.999	0.000			
$q = 1$	Mean	0.158	4.570	0.681	0.019	0.881	93	0.231	−0.003	4.438	0.790	0.004	98	85	187.6
	Std. Err.	0.115	0.491	0.234	0.084	0.091		0.215	0.006	0.590	0.185	0.037			
	Min	−0.156	3.180	0.068	0.000	0.613		−0.343	−0.020	2.785	0.147	0.000			
	Max	0.472	5.949	0.983	0.508	1.102		0.917	0.013	5.817	0.988	0.362			
$q = 2$	Mean	0.035	3.313	0.417	0.124	0.969	73	0.079	−0.002	3.235	0.663	0.005	96	49	32.0
	Std. Err.	0.157	0.683	0.273	0.239	0.116		0.306	0.009	0.770	0.196	0.022			
	Min	−0.322	1.548	0.015	0.000	0.645		−0.786	−0.022	1.217	0.208	0.000			
	Max	0.439	4.965	0.951	0.888	1.200		0.902	0.024	4.866	0.959	0.129			
$q = 3$	Mean	−0.022	2.893	0.453	0.072	1.009	81	0.010	−0.001	2.836	0.667	0.008	98	8	17.3
	Std. Err.	0.158	0.693	0.258	0.181	0.112		0.306	0.010	0.762	0.191	0.062			
	Min	−0.388	1.144	0.020	0.000	0.690		−0.803	−0.024	0.945	0.101	0.000			
	Max	0.389	4.489	0.944	0.846	1.236		0.768	0.023	4.422	0.948	0.619			

N, the number of successful fitting to DAR model from 100 times of random re-sampling of the individual orders.*

of random re-sampling, the PLEC model performed slightly better than the PL model, and species-level modeling slightly better than genus level. Therefore, the PLEC model at the species level performed best among four categories of the models.

(ii) At both genus and species levels, the DAR scaling parameter *z* decreased monotonically with diversity order *q*, and the species level parameters are generally larger than their genus level counterparts. In the case of scaling parameter *z*, larger *z*-value indicates larger PL *slope* or fast change rates of diversity per unit accrual of area. This result should be expected obviously because the differences among individual subjects should be smaller at higher taxonomic level (genus) than lower level (species). In other words, the resolution of higher (genus) taxonomic level is rougher than that of the lower (species) taxonomic level. **Figure 1**

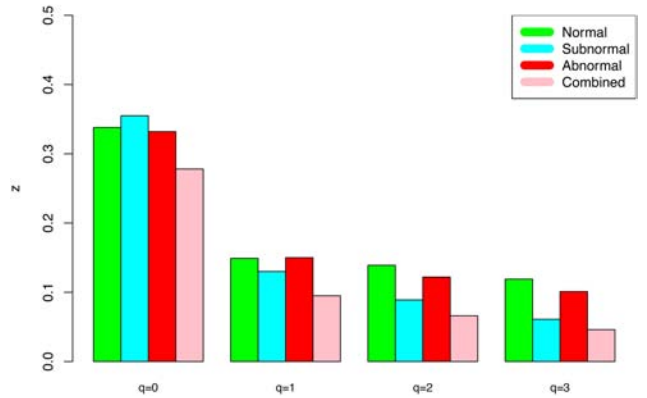


FIGURE 1 | The alpha-DAR profile scaling parameter (*z*-*q* series) for the semen microbiome alpha-diversity at the genus level, for normal, sub-normal, abnormal, and combined groups.

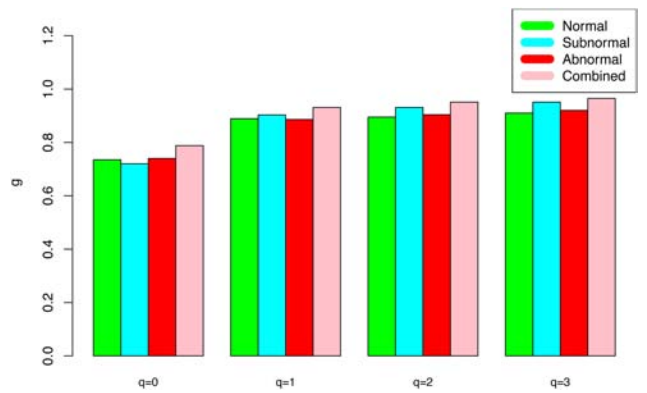


FIGURE 2 | The alpha-PDO profile (*g*-*q* series) for the semen microbiome *alpha*-diversity at the genus level for normal, sub-normal, abnormal, and combined groups, respectively.

exemplified the DAR profiles of the alpha-diversity at the genus level, for the normal, sub-normal, abnormal, and combined groups, respectively.

(iii) At both genus and species levels, the PDO (pair-wise diversity overlap) parameter (g) showed a monotonically increasing trend, which is opposite with that of the scaling parameter (z) as expected. The PDO parameter confirmed the previous finding that semen microbiome has higher similarity (overlap) at genus level than at species level, indicated by higher g-value. **Figure 2** exemplified the PDO profiles of the alpha-diversity at the genus level, for the normal, sub-normal, abnormal, and combined groups, respectively.

(iv) The negative d-values of all PLEC models at both genus and species levels, indicated the existence of asymptote lines and the necessity of introducing more sophisticated PLEC model, which also made the prediction of MAD (maximal accrual diversity) or D_{max} possible. The MAD (D_{max}) decreased with the increase of diversity order (q), as determined by the nature (definition) of the Hill numbers. The MAD at $q = 0$, or ${}^{0}D_{max}$ which is simply the maximal accrual of microbial species (genus) *richness* of the population of individuals. **Figure 3** exemplified the MAD profiles of the alpha-diversity at the genus level, for the normal, sub-normal, abnormal, and combined groups, respectively.

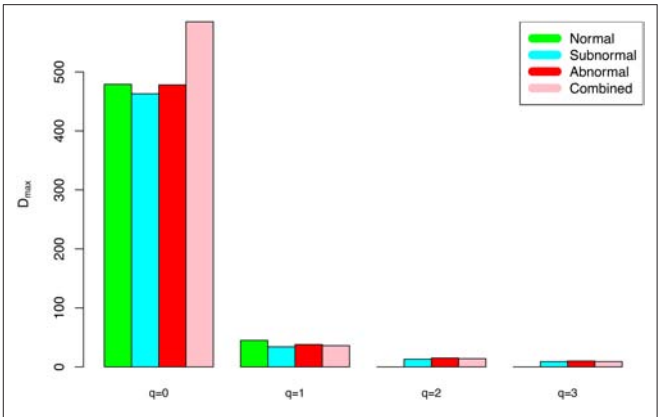


FIGURE 3 | The alpha-MAD profile (D_{max}-q series) for the semen microbiome *alpha*-diversity at the genus level, for normal, sub-normal, abnormal, and combined groups, respectively.

(v) **Table 5** further computed the RIP [Ratio of Individual diversity to Population maximal accrual diversity: Equation (12b)] for all DAR models listed in **Tables 1**–**4**. The left side is the RIP computed from alpha-DAR parameters, and the right side is that computed from beta-DAR parameters. The RIP parameter measures the average level of an individual can represent a population from which he or she comes from. For example, at diversity order $q = 0$, i.e., species (genus) richness level, the alpha-diversity of an individual, on *average*, contains approximately 10.6% (species level) or 29.1% (genus-level) of the diversity accrued by the population. When the diversity order (q) increases, the RIP percentage is also increased, as indicated by **Table 5**. Note that since RIP is defined in terms of an *averaged* individual, it may be a poor representative for a specific individual, especially when the inter-subject heterogeneity of diversity is high. **Figure 4** exemplified the RIP profiles of the alpha-diversity at the genus level, for the normal, sub-normal, abnormal, and combined groups, respectively.

Beta-DAR Modeling

Tables 3, **4** listed the beta-DAR parameters for the human semen microbiome at the genus and species level, respectively. The leftmost column in both the tables are the diversity order ($q = 0, 1, 2, 3$) and, and the parameters for beta-DAR PL models and beta-DAR PLEC models were listed in the left and right side, respectively. From both the tables, we observed the following findings:

(i) The beta-DAR models fitted to the semen microbiome beta-diversity data at both genus and species levels statistically significant ($p < 0.05$ in 7 cases and $p < 0.1$ in 1 case). Judged from the success rates among 100 times of random

TABLE 3 | The parameters of *beta*-DAR (*beta*-diversity area relationship) computed with 100 times of re-sampling at **genus** level.

Diversity order & statistics		Power law (PL)						PL with exponential cutoff (PLEC)							
		z	ln(*c*)	*R*	*p*-value	*g*	*N**	*z*	*d*	ln(*c*)	*R*	*p*-value	*N*	A_{max}	D_{max}
$q = 0$	Mean	0.266	0.294	0.990	0.000	0.797	100	0.338	−0.002	0.152	0.997	0.000	100	140	4.4
	Std. Err.	0.014	0.059	0.005	0.000	0.012		0.025	0.001	0.067	0.002	0.000			
	Min	0.235	0.160	0.975	0.000	0.773		0.272	−0.004	0.023	0.988	0.000			
	Max	0.296	0.424	0.999	0.000	0.823		0.395	−0.001	0.341	0.999	0.000			
$q = 1$	Mean	0.160	0.483	0.800	0.004	0.882	99	0.277	−0.004	0.252	0.895	0.000	100	70	3.2
	Std. Err.	0.061	0.273	0.165	0.045	0.047		0.107	0.004	0.255	0.092	0.000			
	Min	−0.010	−0.103	0.079	0.000	0.772		0.010	−0.012	−0.339	0.483	0.000			
	Max	0.296	1.225	0.982	0.448	1.007		0.556	0.005	0.943	0.984	0.000			
$q = 2$	Mean	0.225	0.595	0.782	0.008	0.829	99	0.393	−0.006	0.264	0.872	0.000	100	69	4.6
	Std. Err.	0.084	0.364	0.166	0.083	0.068		0.171	0.005	0.401	0.107	0.000			
	Min	−0.005	−0.321	0.022	0.000	0.623		−0.066	−0.018	−0.886	0.470	0.000			
	Max	0.462	1.564	0.977	0.829	1.003		0.864	0.009	1.112	0.983	0.000			
$q = 3$	Mean	0.273	0.586	0.804	0.002	0.789	99	0.441	−0.006	0.252	0.874	0.000	100	77	5.6
	Std. Err.	0.092	0.385	0.143	0.015	0.077		0.207	0.006	0.463	0.106	0.000			
	Min	0.033	−0.430	0.150	0.000	0.541		−0.070	−0.022	−1.080	0.509	0.000			
	Max	0.545	1.541	0.980	0.148	0.977		1.016	0.011	1.178	0.985	0.000			

N, the number of successful fitting to DAR model from 100 times of random re–sampling of the individual orders.*

TABLE 4 | The parameters of *beta*-DAR (*beta*-diversity area relationship) computed with 100 times of re-sampling at **species** level.

Diversity order & statistics		Power law (PL)						PL with exponential cutoff (PLEC)							
		z	ln(*c*)	*R*	*p*-value	*g*	*N**	*z*	*d*	ln(*c*)	*R*	*p*-value	*N*	A_{max}	D_{max}
$q = 0$	Mean	0.480	0.389	0.994	0.000	0.605	100	0.588	−0.004	0.174	0.999	0.000	100	161	13.1
	Std. Err.	0.012	0.052	0.003	0.000	0.011		0.028	0.001	0.064	0.001	0.000			
	Min	0.453	0.222	0.985	0.000	0.568		0.534	−0.006	0.003	0.996	0.000			
	Max	0.518	0.506	0.998	0.000	0.631		0.666	−0.002	0.323	1.000	0.000			
$q = 1$	Mean	0.226	0.572	0.884	0.000	0.830	100	0.366	−0.005	0.294	0.944	0.000	100	77	4.6
	Std. Err.	0.051	0.236	0.090	0.000	0.041		0.107	0.004	0.230	0.042	0.000			
	Min	0.090	0.004	0.577	0.000	0.736		0.086	−0.016	−0.241	0.743	0.000			
	Max	0.338	1.205	0.983	0.000	0.935		0.670	0.005	0.816	0.989	0.000			
$q = 2$	Mean	0.235	0.732	0.781	0.000	0.821	100	0.405	−0.006	0.396	0.876	0.000	100	71	5.6
	Std. Err.	0.083	0.377	0.146	0.000	0.068		0.173	0.006	0.360	0.084	0.000			
	Min	0.074	−0.121	0.358	0.000	0.648		−0.028	−0.024	−0.392	0.623	0.000			
	Max	0.435	1.452	0.961	0.000	0.947		0.855	0.011	1.099	0.979	0.000			
$q = 3$	Mean	0.276	0.676	0.765	0.000	0.786	100	0.378	−0.003	0.474	0.838	0.000	100	109	6.5
	Std. Err.	0.103	0.438	0.143	0.002	0.088		0.228	0.008	0.430	0.124	0.001			
	Min	0.057	−0.294	0.248	0.000	0.574		−0.181	−0.024	−0.602	0.317	0.000			
	Max	0.512	1.478	0.945	0.016	0.959		0.868	0.019	1.353	0.980	0.008			

N, the number of successful fitting to DAR model from 100 times of random re-sampling of the individual orders.*

re-sampling, the beta-PLEC model performed slightly better than beta-PL model, and species-level modeling slightly better than the genus-level. Therefore, the beta-PLEC model at the species level performed best among four categories of the models.

(ii) At both genus and species levels, the beta-DAR scaling parameter *z* exhibited a valley-shaped pattern with diversity order (*q*), and the species level parameters are generally larger than their genus level counterparts. In the case of scaling parameter *z*, larger z-value indicates larger slope or faster change rates of diversity per unit change of area accrual. This result should be expected obviously because the differences among individual subjects should be smaller at higher taxonomic level (genus) than lower level (species).

(iii) At both genus and species levels, the beta-PDO parameter (*g*) showed a mountain-shaped trend, which is opposite

TABLE 5 | RIP (ratio of individual diversity to population maximal accrual diversity).

Diversity order (q)	Alpha-RIP (%)		Beta-RIP (%)	
	Genus-level	Species-level	Genus-level	Species-level
$q = 0$	29.1	10.6	30.5	11.3
$q = 1$	66.0	51.5	50.7	38.5
$q = 2$	73.4	85.8	39.4	37.1
$q = 3$	79.3	100	23.5	39.2

with that of the scaling parameter (z) as expected. The beta-PDO parameter confirmed previous finding that semen microbiome has higher level of similarity at genus level than at species level, indicated by higher g-value, which measures the pair-wise diversity overlap (similarity).

(iv) The negative d-values of all beta-PLEC models at both genus and species levels, indicated the existence of asymptote lines and the necessity for introducing the more sophisticated beta-PLEC model, which also made the prediction of beta-D_{max} (beta-MAD) possible. The beta-MAD-q or beta-D_{max}-q, or beta-MAD profile, exhibited a valley-shaped trend, which is consistent with the z-q series or DAR profile. The beta-MAD or at $q = 0$, ${}^0D_{max}$, which is simply the species (genus) *richness* or the number of species (genus) in the cohort (or population).

(v) **Table 5** listed the RIP profile, i.e., the ratio of individual diversity to population maximal accrual diversity (Equation 12b), at different diversity order (q), for all DAR models listed in **Tables 1**–**4**. The two right columns were computed for beta-DAR models from **Tables 3**, **4** (the left side for alpha-DAR). The RIP parameter (profile) measures the average level of an individual can represent a population from which he or she comes from. For example, when q = 0 (species or genus richness level), beta-RIP = 11.3% at species level, and beta-RIP = 30.5% at genus level. This suggested that beta-diversity of an individual, on average, contains ∼11.3% (at the species level) and 30.5% (at the genus level) of the diversity accrued at the population level. When the diversity order (q) increases, the RIP also increased accordingly.

An interesting observation is that alpha-RIP profile and beta-RIP profile exhibited different patterns: the former is monotonically increasing, but the latter is mountain-shaped. This pattern is clear from comparing of the left side and right side of **Table 5**.

DISCUSSION

The results of DAR analysis presented above revealed that fertility status (normal, subnormal, abnormal) did not have a significant influence on biogeography of semen microbiome, specifically, on the inter-subject (spatial) heterogeneity in terms of either alpha-diversity or beta-diversity. Previous studies have suggested changes in semen microbiome diversity associated with fertility health (Hou et al., 2013; Weng et al., 2014), although no rigorous statistical tests were performed with the published

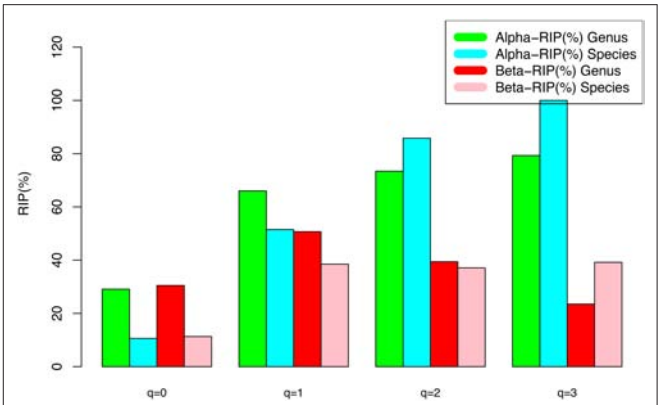


FIGURE 4 | The RIP-profile (RIP-q series) for the semen microbiome diversity (alpha and beta diversity, respectively) at the genus level, for the normal, sub-normal, abnormal, and combined groups, respectively.

studies. Furthermore, the diversity of a microbiome sample *per se* and the diversity scaling (or spatial heterogeneity changes, a topic of this study) within a population are very different concepts. Logically, the change of individual diversity does not necessary lead to changes of the diversity heterogeneity among individuals. Therefore, the lack of differences in the diversity scaling parameter (z) and other DAR parameters, among three groups with different fertility status do not contradict the published studies on the human semen microbiome.

The lack of significant differences among various fertility groups actually simplified our study, enabled us to build the DAR models for a general Chinese population. Using the DAR models, we were able to (i) estimate the diversity changes of semen microbiome in a human cohort (population) or DAR profile; (ii) predict the maximal accrual diversity (MAD) of semen microbiome in a human cohort (population) or the MAD profile; (iii) estimate the PDO (pair-wise diversity overlap or similarity) between two individuals or the PDO profile; (iv) assess the RIP profile (i.e., the ratio of individual diversity to population accrual diversity), which measures the level an individual can represent a population which he belongs to. The "*profiles*" provide series of key parameter associated with different diversity order (q), which weights diversity differently: from species richness ($q = 0$), where all species are weighted equally, to $q = 3$, where dominant species were weighted for more and rare species were weighted for less. These parameters sketched out the biogeography "maps" of the human semen microbiome in terms of the four profiles: the DAR-, PDO-, MAD-, and RIP profiles. Together, the four profiles (maps) comprehensively sketched out the biogeography of semen microbiome—the spatial distribution or inter-subject heterogeneity of semen microbiome diversity at different diversity orders (q). The different biogeography maps are similar to different geography maps, each may with different utilization (e.g., rainfall map vs. biodiversity map, both of different utilizations). Using another analogy, maps at different diversity order (q) are similar to the maps with different scales or resolutions.

Hence, similar to the obvious significance of geographic maps, our biogeographic maps for the human semen

microbiome diversity distribution should be rather important for further investigating the spatial distribution (or inter-subject heterogeneity) of the semen microbiome and their biomedical implications. A limitation of this study is that the datasets we used were limited to a Chinese population. We hope that future studies will include datasets from other ethnic groups.

AUTHOR CONTRIBUTIONS

ZM designed the study, interpreted the results and wrote the paper. LL performed the computation and participated the interpretation of the results.

FUNDING

This study received funding from the following sources: National Science Foundation of China (Grant No. 71473243), Cloud-Ridge Industry Technology Leader Grant, A China-US International Cooperation Project on Genomics/Metagenomics Big Data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.03333/full#supplementary-material

REFERENCES

Akutsu, T., Motani, H., Watanabe, K., Iwase, H., and Sakurada, K. (2012). Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid. *Leg Med.* 14, 160–162. doi: 10.1016/j.legalmed.2012.01.005

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2′s q2-feature-classifier plugin. *Microbiome* 6:90. doi: 10.1186/s40168-018-0470-z

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, 335–336. doi: 10.1038/nmeth.f.303

Chao, A., Chiu, C. H., and Hsieh, T. C. (2012). Proposing a resolution to debates on diversity partitioning. *Ecology* 93, 2037–2051. doi: 10.1890/11-1817.1

Chao, A., Chiu, C. H., and Jost, L. (2014). Unifying species diversity, phylogenetic diversity, functional diversity and related similarity and differentiation measures through Hill numbers. *Annu. Rev. Ecol. Evol. Syst.* 45, 297–324. doi: 10.1146/annurev-ecolsys-120213-091540

Connor, E. F., and McCoy, E. D. (1979). The statistics and biology of the species–area relationship. *Am. Nat.* 113, 791–833. doi: 10.1086/283438

De Francesco, M. A., Negrini, R., Ravizzola, G., Galli, P., and Manca, N. (2011). Bacterial species present in the lower male genital tract: a five-year retrospective study. *Eur. J. Contracept. Reprod. Health Care* 16, 47–53. doi: 10.3109/13625187.2010.533219

Diemer, T., Huwe, P., Ludwig, M., Hauck, E. W., and Weidner, W. (2003). Urogenital infection and sperm motility. *Andrologia* 35, 283–287. doi: 10.1111/j.1439-0272.2003.tb00858.x

Domes, T., Lo, K. C., Grober, E. D., Mullen, J. B., Mazzulli, T., and Jarvi, K. (2012). The incidence and effect of bacteriospermia and elevated seminal leukocytes on semen parameters. *Fertil. Steril.* 97, 1050–1055. doi: 10.1016/j.fertnstert.2012.01.124

Drakare, S., Lennon, J. J., and Hillebrand, H. (2006). The imprint of the geographical, evolutionary and ecological context on species–area relationships. *Ecol. Lett.* 9, 215–227 doi: 10.1111/j.1461-0248.2005.00848.x

Harte, J., Kinzig, A., and Green, J. (1999). Self-similarity in the distribution and abundance of species. *Science* 284, 334–336. doi: 10.1126/science.284.5412.334

He, F., and Hubbell, S. P. (2011). Species-area relationships always overestimate extinction rates from habitat loss. *Nature* 473, 368–371. doi: 10.1038/nature09985

Helmus, M. R., Mahler, D. L., and Losos, J. B. (2014). Island biogeography of the Anthropocene. *Nature* 513, 543–547 doi: 10.1038/nature13739

Henkel, R., Ludwig, M., Schuppe, H. C., Diemer, T., Schill, W. B., and Weidner, W. (2006). Chronic pelvic pain syndrome/chronic prostatitis affect the acrosome reaction in human spermatozoa. *World J. Urol.* 24, 39–44. doi: 10.1007/s00345-005-0038-y

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54, 427–432. doi: 10.2307/1934352

Hou, D., Zhou, X., Zhong, X., Settles, M. L., Herring, J., Wang, L., et al. (2013). Microbiota of the seminal fluid from healthy and infertile men. *Fertil. Steril.* 100, 1261–1269. doi: 10.1016/j.fertnstert.2013.07.1991

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology* 88, 2427–2439. doi: 10.1890/06-1736.1

Keck, C., Gerber-Schäfer, C., Clad, A., Wilhelm, C., and Breckwoldt, M. (1998). Seminal tract infections: impact on male fertility and treatment options. *Hum. Reprod. Update* 4, 891–903. doi: 10.1093/humupd/4.6.891

Kiessling, A. A., Desmarais, B. M., Yin, H. Z., Loverde, J., and Eyre, R. C. (2008). Detection and identification of bacterial DNA in semen. *Fertil. Steril.* 90, 1744–1756. doi: 10.1016/j.fertnstert.2007.08.083

Lbadin, O. K., and Ibeh, I. N. (2008). Bacteriospermia and sperm quality in infertile male patient at University of Benin Teaching Hospital, Benin City, Nigeria. *Malaysian J. Microbiol.* 4, 65–67. doi: 10.21161/mjm.12108

Lomolino, M. V. (2000). Ecology's most general, yet protean pattern: the species–area relationship. *J. Biogeogr.* 27, 17–26. doi: 10.1046/j.1365-2699.2000.00377.x

Ma, Z. S. (2017). "Measuring Microbiome Diversity and Similarity with Hill Numbers" in *Metagenomics Perspectives, Methods, and Applications 2018*, ed M. Nagarajan (Elsevier Inc.), 157–178.

Ma, Z. S. (2018a). Extending species-area relationships (SAR) to diversity-area relationships (DAR). *Ecol. Evol.* 8, 10023–10038. doi: 10.1002/ece3.4425

Ma, Z. S. (2018b). Sketching the human microbiome biogeography with DAR (Diversity Area Relationship) profiles. *Microb. Ecol.* doi: 10.1007/s00248-018-1245-6. [Epub ahead of print].

MacArthur, R. H., and Wilson, E. O. (1967). *The Theory of Islands Biogeography*. Princeton, NJ: Princeton University Press.

Moretti, E., Capitani, S., Figura, N., Pammolli, A., Federico, M. G., Giannerini, V., et al. (2009). The presence of bacteria species in semen and sperm quality. *J. Assist. Reprod. Genet.* 26, 47–56. doi: 10.1007/s10815-008-9283-5

Ochsendorf, F. R. (2008). Sexually transmitted infections: impact on male fertility. *Andrologia* 40, 72–75. doi: 10.1111/j.1439-0272.2007.00825.x

Plotkin, J. B., Potts, M. D., Yu, D. W., Bunyavejchewin, S., Condit, R., Foster, R., et al. (2000). Predicting species diversity in tropical forests. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10850–10854. doi: 10.1073/pnas.97.20.10850

Preston, F. W. (1960). Time and space and the variation of species. *Ecology* 41, 611–627. doi: 10.2307/1931793

Renyi, A. (1961). "On measures of information and entropy," in *The 4-th Berkeley Symposium on Mathematics, Statistics and Probability*, ed A. Neyman (Berkeley, CA), 547–561.

Rosenzweig, M. L. (1995). *Species Diversity in Space and Time*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511623387

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

Sizling, A. L., Kunin, W. E., Sizlingová, E., Reif, J., and Storch, D. (2011). Between geometry and biology: the problem if universality of the species-area relationship. *Am. Nat.* 178, 602–611. doi: 10.1086/662176

Storch, D., Keil, P., and Jetz, W. (2012). Universal species–area and endemics–area relationships at continental scales. *Nature* 488, 78–83. doi: 10.1038/nature11226

Tjørve, E. (2009). Shapes and functions of species–area curves (II): a review of new models and parameterizations. *J. Biogeogr.* 36, 1435–1445. doi: 10.1111/j.1365-2699.2009.02101.x

Tjørve, E., and Tjørve, K. M. (2008). The species–area relationship, self-similarity, and the true meaning of the z-value. *Ecology* 89, 3528–3533. doi: 10.1890/07-1685.1

Triantis, K. A., Guilhaumon, F., and Whittaker, R. J. (2012). The island species–area relationship: biology and statistics. *J. Biogeogr.* 39, 215–231 doi: 10.1111/j.1365-2699.2011.02652.x

Ulrich, W., and Buszko, J. (2003). Self-similarity and the species–area relation of Polish butterflies. *Basic Appl. Ecol.* 4, 263–270. doi: 10.1078/1439-1791-00139

Watson, H. C. (1835). *Remarks on Geographic Distribution of British Plants*. London: Longman, Rees, Orme, Brown, Green and Longman.

Weng, S.-L., Chiu, C.-M., Lin, F.-M., Huang, W.-C., Liang, C., Yang, T, et al. (2014). Bacterial communities in semen from men of infertile couples: metagenomic sequencing reveals relationships of seminal microbiota to semen quality. *PLoS ONE* 9:e110152. doi: 10.1371/journal.pone.0110152

Whittaker, R. J., and Triantis, K. A. (2012). The species–area relationship: an exploration of that 'most general, yet protean pattern'. *J. Biogeogr.* 39, 623–626. doi: 10.1111/j.1365-2699.2012.02692.x

Conflict of Interest Statement: 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.

Copyright © 2019 Ma and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 14 February 2019
doi: 10.3389/fmicb.2019.00232

Dynamic Development of Fecal Microbiome During the Progression of Diabetes Mellitus in Zucker Diabetic Fatty Rats

Wen Zhou[1], Huiying Xu[1], Libin Zhan[1], Xiaoguang Lu[2]* and Lijing Zhang[1]*

[1] *Modern Research Laboratory of Spleen Visceral Manifestations Theory, Basic Medical College, Nanjing University of Chinese Medicine, Nanjing, China,* [2] *Department of Emergency Medicine, Zhongshan Hospital, Dalian University, Dalian, China*

Background: Although substantial efforts have been made to link the gut microbiota to type 2 diabetes, dynamic changes in the fecal microbiome under the pathological conditions of diabetes have not been investigated.

Methods: Four male Zucker diabetic fatty (ZDF) rats received Purina 5008 chow [protein = 23.6%, Nitrogen-Free Extract (by difference) = 50.3%, fiber (crude) = 3.3%, ash = 6.1%, fat (ether extract) = 6.7%, and fat (acid hydrolysis) = 8.1%] for 8 weeks. A total of 32 stool samples were collected from weeks 8 to 15 in four rats. To decipher the microbial populations in these samples, we used a 16S rRNA gene sequencing approach.

Results: Microbiome analysis showed that the changes in the fecal microbiome were associated with age and disease progression. In all the stages from 8 to 15 weeks, phyla *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria* primarily dominated the fecal microbiome of the rats. Although *Lactobacillus* and *Turicibacter* were the predominant genera in 8- to 10-week-old rats, *Bifidobacterium*, *Lactobacillus*, *Ruminococcus*, and *Allobaculum* were the most abundant genera in 15-week-old rats. Of interest, compared to the earlier weeks, relatively greater diversity (at the genus level) was observed at 10 weeks of age. Although the microbiome of 12-week-old rats had the highest diversity, the diversity in 13–15-week-old rats was reduced. Spearman's correlation analysis showed that F/B was negatively correlated with age. Random blood glucose was negatively correlated with *Lactobacillus* and *Turicibacter* but positively correlated with *Ruminococcus* and *Allobaculum* and Simpson's diversity index.

Conclusion: We demonstrated the time-dependent alterations of the abundance and diversity of the fecal microbiome during the progression of diabetes in ZDF rats. At the genus level, dynamic changes were observed. We believe that this work will enhance our understanding of fecal microbiome development in ZDF rats and help to further analyze the role of the microbiome in metabolic diseases. Furthermore, our work may also provide an effective strategy for the clinical treatment of diabetes through microbial intervention.

Keywords: 16S gene sequencing, fecal microbiome, type 2 diabetes mellitus, gut microbiota, time series, rat microbiome

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Jia Lianqun,
Liaoning University of Traditional Chinese Medicine, China
Jin-Rong Zhou,
Harvard Medical School, United States

****Correspondence:***
Libin Zhan
zlbnj@njucm.edu.cn
Xiaoguang Lu
dllxg@126.com

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

Received: *07 November 2018*
Accepted: *28 January 2019*
Published: *14 February 2019*

Citation:
Zhou W, Xu H, Zhan L, Lu X and Zhang L (2019) Dynamic Development of Fecal Microbiome During the Progression of Diabetes Mellitus in Zucker Diabetic Fatty Rats. Front. Microbiol. 10:232. doi: 10.3389/fmicb.2019.00232

INTRODUCTION

Type 2 diabetes mellitus (T2DM) is currently the most prevalent metabolic disease in the world and is characterized by insulin resistance, with an initial increase in insulin secretion, but subsequent beta cell death and insulin insufficiency over time. According to the International Diabetes Federation, T2DM will affect 693 million people worldwide by 2045 (Cho et al., 2018). T2DM is a multifactorial disorder, with pathogenic contributions from genetic, environmental, and lifestyle factors (Mengual et al., 2010; Saxena et al., 2012). The gut microbiota has increasingly been recognized as a key contributor to T2DM, and T2DM can be linked to dysbiosis of the intestinal microbiota (Cox et al., 2014; Forslund et al., 2015; Yano et al., 2015). Two independent studies based on fecal samples from European and Chinese populations showed increased abundances of opportunistically pathogenic *Clostridium* species and decreased abundances of butyrate-producing *Roseburia*, *Faecalibacterium*, and *Eubacterium* species associated with T2DM patients (Qin et al., 2012; Karlsson et al., 2013). Karlsson et al. (2013) also found that increased abundances of *Lactobacillus gasseri* and *Streptococcus mutans* can predict insulin resistance, while Qin et al. (2012) found enrichment in *Escherichia coli* associated with current T2DM patients. Some studies have also found that pro-inflammatory bacteria such as *Ruminococcus gnavus* and *Bacteroides* spp. are more common in the feces of T2DM patients (Everard and Cani, 2013). Numerous studies have shown significant changes in the composition and diversity of the fecal microflora under conditions of diabetes. Studies also speculate that changes in the composition and diversity of feces can determine the prognosis and severity of T2DM. Our understanding of the relevance of the microbiome in metabolic diseases might be enhanced by systematically assessing the role of the fecal microbiota in disease performance and its control.

An understanding of the fecal microbiome in T2DM has recently arisen by analyzing microbial populations found in fecal samples at a certain point in time. Although such assessments of the fecal microbiome composition and diversity in T2DM are valuable, they are time-limited and do not reflect the dynamic changes of microbial flora in the progression of T2DM. Several studies have reported that the fecal microbiome differs at different times during the progression of T2DM (Horie et al., 2017; Liu et al., 2017). Therefore, more work needs to be done to determine the role of the fecal microbiome diversity and composition and their association with T2DM. Due to ethical issues and the availability of a limited number of samples, analysis of the fecal microbiome and its role in the disease pathogenesis of diabetes in humans is limited. Thus, to establish the diabetic fecal microbiome, small animal models can be used. In these models, fecal samples can be conveniently collected, thereby allowing for investigation of the microbiome contribution in T2DM. In fact, to understand the role of the microbiome in T2DM, many animal models have been widely used (Bagarolli et al., 2017; Bindels et al., 2017; Caparros-Martin et al., 2017). In addition, evidence emerging from animal models shows that many of the symptoms associated with diabetic syndrome and insulin sensitivity may be improved through replenishing probiotics (*Lactobacillus rhamnosus*, *Lactobacillus acidophilus*, and *Bifidobacterium*) and butyric-acid producing bacteria *Clostridium butyricum* (Bagarolli et al., 2017; Jia et al., 2017). Although some studies have used rat models to elucidate the microbiome's role in T2DM (Goldsmith et al., 2017; Kim et al., 2017), in the field of T2DM, one of the major unanswered questions is whether the microbiome can be utilized to alleviate diabetic pathologies.

Currently, only a few studies have examined details about the compositional dynamics of the diabetic microbiome (Horie et al., 2017; Liu et al., 2017). Since most of these studies were conducted at some point in the course of T2DM development, they do not provide an insight into the development of the diabetic fecal microbiome. Using animal models might establish a better understanding of the fecal microbiome in the progression of T2DM, and such knowledge can enhance our understanding of the microbiome effects on T2DM. ZDF rats with a missense mutation (fatty, fa) in the leptin receptor gene can develop obesity, insulin resistance, and T2DM (Phillips et al., 1996; Yamashita et al., 1997; Da Silva et al., 1998; Yokoi et al., 2013). Male ZDF rats exhibit an age-dependent diabetic phenotype that develops hyperglycemia at 8 weeks of age and the blood glucose level remains high throughout its lifespan (De Lemos et al., 2007). Due to these characteristics, ZDF rats are an attractive experimental model for this study. In this study, we monitored body weight, food intake, water intake, rectal temperature, RBG, OGTT, and the fecal microbiome from 8 to 15 weeks of age in ZDF rats. We analyzed the fecal microbiome at different time points in diabetic rats and tracked changes in microbial diversity. A deep sequencing of 16S rRNA genes amplified from genomic DNA isolated from the rat feces was used. To this end, we also performed non-parametric Spearman's correlation analysis to evaluate associations between physiological characteristics and the microbiome in ZDF rats.

MATERIALS AND METHODS

Experimental Design

This study was done longitudinally and its primary purpose was to understand the changes in fecal microbiome composition during diabetes progression in four ZDF rats. We studied the microbiome from week 8 onward to week 15 at 1-week intervals. Studies were performed using ZDF rats as they have been shown to exhibit hyperinsulinemia and hyperglycemia (De Lemos et al., 2007) and are thus a good model of T2DM.

Ethics Statement

In the present study, the animal experiments used rats and were approved by the Animal Ethics Committee of Nanjing University of Chinese Medicine (Approval No. ACU170606). All animal experiments were conducted in accordance with the National Institutes of Health Guide for the Care and Use of

Abbreviations: F/B, *Firmicutes*/*Bacteroidetes*; IQR, interquartile range; OGTT, oral glucose tolerance test; OTUs, operational taxonomic units; PCoA, principal coordinates analysis; RBG, random blood glucose; SDI, Simpson's diversity index; T2DM, type 2 diabetes mellitus; ZDF, Zucker diabetic fatty.

Laboratory Animals at Nanjing University of Chinese Medicine (Nanjing, China).

Animal

Four male 6-week-old ZDF rats were purchased from Vital River Laboratories (Beijing, China) and housed in a specific pathogen-free animal experimental center in Nanjing University of Chinese Medicine. Animals were fed autoclaved Purina 5008 chow [protein = 23.6%, Nitrogen-Free Extract (by difference) = 50.3%, fiber (crude) = 3.3%, ash = 6.1%, fat (ether extract) = 6.7%, and fat (acid hydrolysis) = 8.1%; Vital River Laboratories, Beijing, China], had free access to autoclaved water, and housed at 24°C ± 2°C, humidity 65% ± 5%, with a 12 h light-dark cycle. During the trial, body weight, food and water intake, and rectal temperature were measured daily. All rats were in one group and housed in one cage during the study.

Random Blood Glucose Test

Random blood glucose was measured weekly to examine the progression of diabetes in ZDF rats. Glucose levels in tail blood samples were measured from weeks 8 to 15 using a glucometer (CareSens, I-SENS, Anyang, South Korea). The rats were not fasted for RBG tests.

Oral Glucose Tolerance Test

Zucker diabetic fatty rats were fasted for 14 h (overnight) and then the OGTT was performed with a glucose solution in saline at 2 g/kg. Tail blood was sampled at 0, 30, 60, and 120 min after glucose administration. Glucose levels were determined immediately with a glucometer (CareSens, I-SENS).

Stool Sample Collection and DNA Extraction

One fresh fecal sample was collected directly from the anus into a sterile tube from each rat weekly, avoiding contact with rat skin or urine (see **Supplementary Table S1**). A total of 32 stool samples were collected from weeks 8 to 15 in four ZDF rats and stored at −80°C prior to processing. Bacterial DNA was extracted from feces using the Fast DNA SPIN extraction kit (MP Biomedicals, Santa Ana, CA, United States) according to the manufacturer's instructions and stored at −20°C before further analysis. The quantity and quality of extracted DNA were measured using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, United States) and agarose gel electrophoresis, respectively.

16S rRNA Amplification and Sequencing

PCR amplification of the bacterial 16S rRNA genes (V3–V4 region) was carried out using forward primer 338F (5′-ACTCCTACGGGAGGCAGCA-3′) and reverse primer 806R (5′-GGACTACHVGGGTWTCTAAT-3′). Sample-specific 7-bp barcodes were incorporated into the primers for multiplex sequencing. PCR components contained 5 μl Q5 reaction buffer (5×), 5 μl Q5 High-Fidelity GC buffer (5×), 0.25 μl Q5 High-Fidelity DNA polymerase (5 U/μl), 2 μl dNTPs (2.5 mM), 1 μl of each forward and reverse primers (10 μM), 2 μl DNA template, and 8.75 μl ddH_2O. Thermal cycling included initial denaturation for 2 min at 98°C, followed by 25 cycles including denaturation for 15 s at 98°C, annealing for 30 s at 55°C, and extension for 30 s at 72°C, and a final extension of 5 min at 72°C. PCR amplicons were purified using Agencourt AMPure Beads (Beckman Coulter, Indianapolis, IN, United States) and quantified with the PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA, United States). After the individual quantification step, amplicons were combined in equal amounts and subjected to 2 × 300 bp sequencing of the end using the Illumina MiSeq platform and the MiSeq kit v3 from Shanghai Personal Biotechnology Co., Ltd. (Shanghai, China). Sequencing data were processed using a quantitative analysis of microbial ecology (QIIME, v1.8.0). In brief, original sequencing reads that perfectly matched the barcode were assigned to the corresponding samples and identified as valid sequences. Low-quality sequences (Gill et al., 2006; Chen and Jiang, 2014) were filtered by the following criteria: sequences <150 bp in length, sequences with average Phred scores <20, sequences containing indefinite bases, and sequences containing single nucleotide repeats of >8 bp. Paired-end reads were assembled using FLASH (Magoc and Salzberg, 2011). After chimera detection, the remaining high-quality sequences were clustered into OTUs with 97% sequence identity by UCLUST (Edgar, 2010). The default parameters were used to select the representative sequence from each OTU. Using the best hits (Altschul et al., 1997), OTU taxonomy classification was performed by a BLAST search on the representative set of sequences against the Greengenes database (Desantis et al., 2006). The abundance of each OTU in each sample and the taxonomy of these OTUs were recorded by generating an OTU table. OTUs with a total content of less than 0.001% in all samples were discarded. To minimize the difference in the depth of sequencing across samples, the average analysis of 100 evenly resampled OTU subsets under the 90% of the minimum sequencing depth was performed to generate an average, rounded dilution OTU table.

Bioinformatics Analysis

Sequencing data were evaluated using the QIIME and R software packages (v3.2.0). The OTU table in QIIME was used to calculate the α diversity index of the OTU level, such as the Shannon diversity index and the SDI. Principal weighted UniFrac distance metrics (Lozupone and Knight, 2005) were used for principal coordinate analysis (PCoA). Diversity was assessed using the Simpson Diversity Index (SDI) by calculating "inverse" (1/λ) and "complement" (1-λ) SDI. Higher SDI values indicated higher microbial diversity. Based on the occurrence of OTUs across samples, a petal diagram was created to visualize the shared and unique OTUs among samples or groups by the R package "Venn Diagram." Metastats (White et al., 2009) was used to statistically compare the abundance of taxa at the level of phylum and genus among samples or groups.

Statistical Analysis

The physiological characteristics data of the ZDF rats are presented as mean ± SD. Statistical analyses among different ages were performed by repeated ANOVA, followed by Tukey's

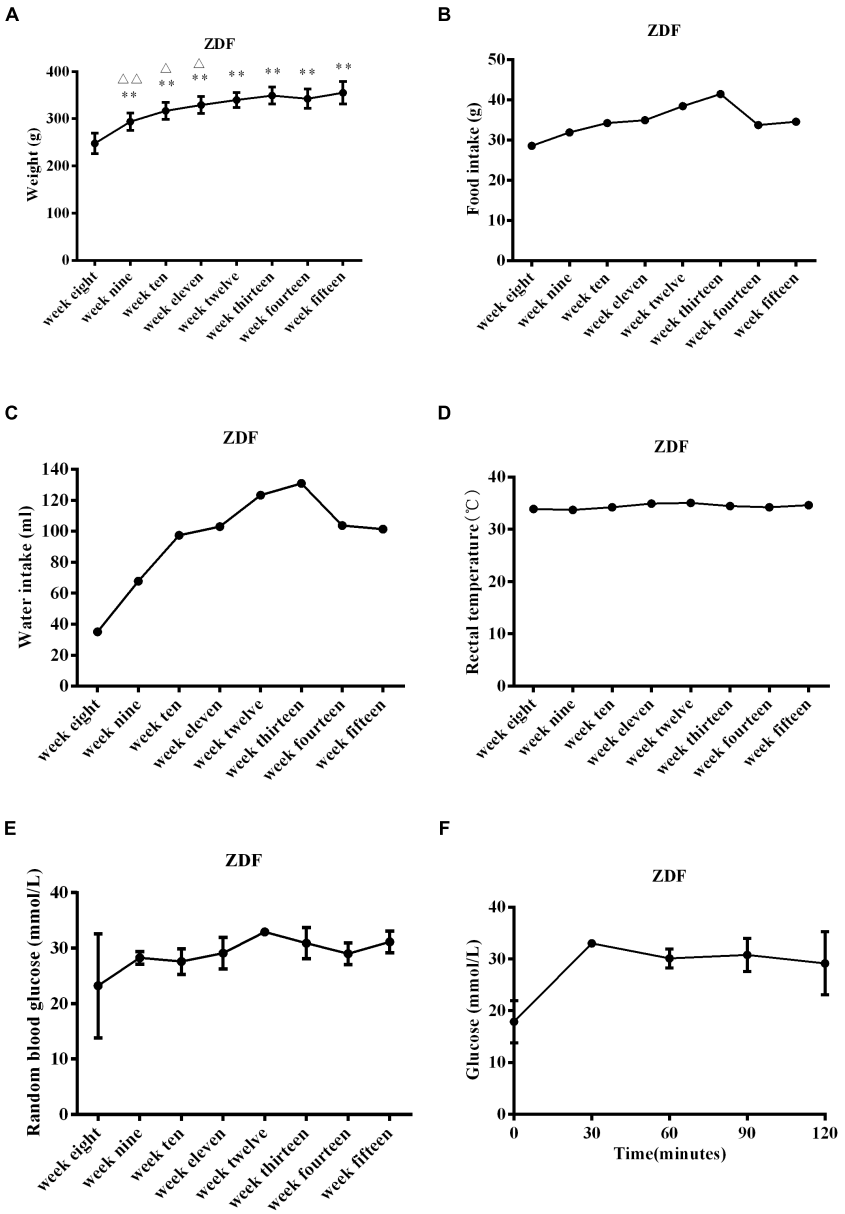


FIGURE 1 | Physiological characteristics in ZDF rats ranging from 8 to 15 weeks old. **(A)** Body weight gain in ZDF rats. **(B)** Food intake changes in ZDF rats. **(C)** Water intake changes in ZDF rats. **(D)** Rectal temperature changes in ZDF rats. **(E)** Random blood glucose levels in ZDF rats. **(F)** Oral glucose tolerance test measured at week 14 in ZDF rats. Data are expressed as mean ± SEM. $N = 4$ in individual groups; data were analyzed by repeated ANOVA: compared with week 8, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; compared with the previous week, $^{\Delta}p < 0.05$, $^{\Delta\Delta}p < 0.01$, $^{\Delta\Delta\Delta}p < 0.001$.

honestly significant difference or Dunnett's *post hoc* test with SPSS 19.0 (IBM, Chicago, IL, United States), considering a P-value ≤ 0.05 as statistically significant. Correlations between physiological characteristics data and either F/B ratio or genus were tested by Spearman's correlation analysis using Prism 5 (GraphPad, La Jolla, CA, United States).

Sequence Accession Numbers

The datasets generated in this study are available through the NCBI Sequence Read Archive (accession number SRP148630).

RESULTS

Physiological Characteristics of ZDF Rats From 8 to 15 Weeks

Zucker diabetic fatty rats gained significantly more weight from 9 to 15 weeks of age compared to weights at 8 weeks of age ($P < 0.01$ at weeks 9–15, **Figure 1A**). Compared with the previous week, ZDF rats gained significantly more weight at 9, 10, and 11 weeks of age ($P < 0.01$ at week 9, $P < 0.05$ at weeks 10–11, **Figure 1A**). ZDF rats generally experienced an upward trend

in food and water intake from 8 to 15 weeks (**Figures 1B,C**), but the rectal temperature remained stable (**Figure 1D**). Insulin sensitivity was assessed by measuring RBG levels and by the OGTT at week 14. RBG levels in 8-week-old ZDF rats reached diabetes status (**Figure 1E**). The OGTT showed that the blood glucose level reached the highest value at 30 min, and then gradually decreased, but still could not recover to the initial value at 120 min (**Figure 1F**). These findings are consistent with previous reports (Jourdan et al., 2013; Wessels et al., 2015; Van Bree et al., 2016; Szokol et al., 2017) and indicate that the ZDF rats presented with pathological conditions of diabetes. The disease was generally aggravated with age, glucose tolerance was impaired, and insulin sensitivity was reduced.

Developing T2DM Harbors Temporally Dynamic Microbial Diversity

The progression of diabetes may be associated with microbiome dynamic changes; thus, we tracked the fecal microbiome changes in rats from 8 weeks of age until 15 weeks of age. Of note, we did not include rats older than 15 weeks of age in this work because published literature suggests that male ZDF rats exhibit significant diabetic complications at 15 weeks of age (Gu et al., 2017). A total of 1,944,426 16S rRNA (V3–V4 region) reads were obtained, averaging 60,763 reads per sample. Reads were undertaken to generate a total of 44,613 OTUs, which could be further grouped into ∼315 unique OTUs. Collectively, these sequences represented 247 unique genera. The average Shannon Diversity Index for all time points ranged from 4.87 to 6.28, with an average of 5.52 (confidence intervals for all SDI values are provided in **Supplementary Table S2**). Between Shannon and Simpson's diversity indices, there was a consistent trend. Using the SDI could clearly visualize the trends (**Figure 2**). SDI described an increase in diversity from 9 to 12 weeks in ZDF rats, with the highest diversity observed at 12 weeks of age, followed by a slight decrease at 13–15 weeks of age. This trend was repeatable using the inverse SDI (**Supplementary Figure S1A**). The median and inter-quartile range (IQR) are provided in **Supplementary Figure S1B**.

The cluster heatmap for each genus per week is shown in **Figure 3**. The abundance levels of each genus in the cluster heatmap revealed the weekly dominant genera. Overall, the fecal microbiome consisted of unique genera that can reflect the diversity and dynamic changes of a microbial population.

Identification of Core Microbial Communities in the Diabetic Stage of ZDF Rats

The number of rats per week was 4 (see **Supplementary Table S3** for the number of each sample). We plotted the weighted UniFrac distances for all weeks (**Figure 4**) to compare abundance across weeks. Inter-week weighted UniFrac distances were longer than intra-week weighted UniFrac distances.

At the phyla level, compared to the relative percent abundance, more than 90% of the microbial population in ZDF rats from weeks 8 to 15 consisted of the phyla *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria*. During 8–15 weeks of age, the most abundant phylum was *Firmicutes* (**Figure 5**). At 8–9 weeks of age, the predominant phyla were *Firmicutes* and *Bacteroidetes*. *Actinobacteria* gradually increased from the 10th week of age until the 15th week of age. *Proteobacteria* increased significantly in ZDF rats at 15 weeks of age compared to other ages. We have plotted the mean abundance measure along with the standard error for individual phyla (**Supplementary Figure S2**). It is worth noting that the percent abundance of different phyla varied at every week, thereby suggesting a dynamic microbial ecosystem in ZDF rats.

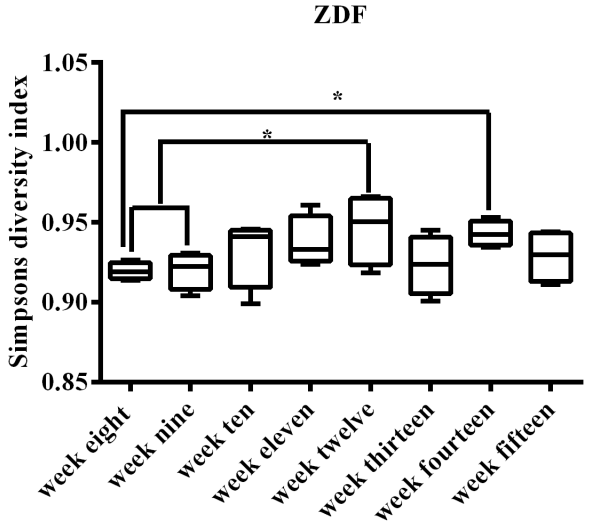


FIGURE 2 | Simpson's diversity index for all 32 samples representing the weeks of development T2DM. Early weeks (8–9) have lower diversity, and the increase in diversity at 10 weeks of age can be clearly visualized from the box plots. Median (line within the box) and minimum and maximum values (whiskers) are illustrated by box and whisker plots. Compared within the weeks, $*p < 0.05$.

Grouping of Microbial Abundance in the Feces of ZDF Rats Shows Temporal Signatures

To identify differences and similarities between the microbial populations in different samples, cluster analyses based on weighted UniFrac distances (Lozupone and Knight, 2005) were carried out. These analyses revealed that weeks 8–10 and 11–13 showed mixed effects and formed two distinct clusters (**Figure 6**). Some samples clustered with other samples from the same week, thus exhibiting high specificity (samples from week 12). Samples from other weeks either clustered non-specifically with other samples or clustered with the nearest neighboring time point (weeks 11–12 and 13–14).

To visualize whether the samples could form distinct clusters, weighted UniFrac distances were used for the principal coordinate analysis (PCoA). Whereas samples from week 8 (red circle), week 9 (blue circle), and week 10 (brown circle) grouped together in a cluster (along the PC3 axis), the remaining samples

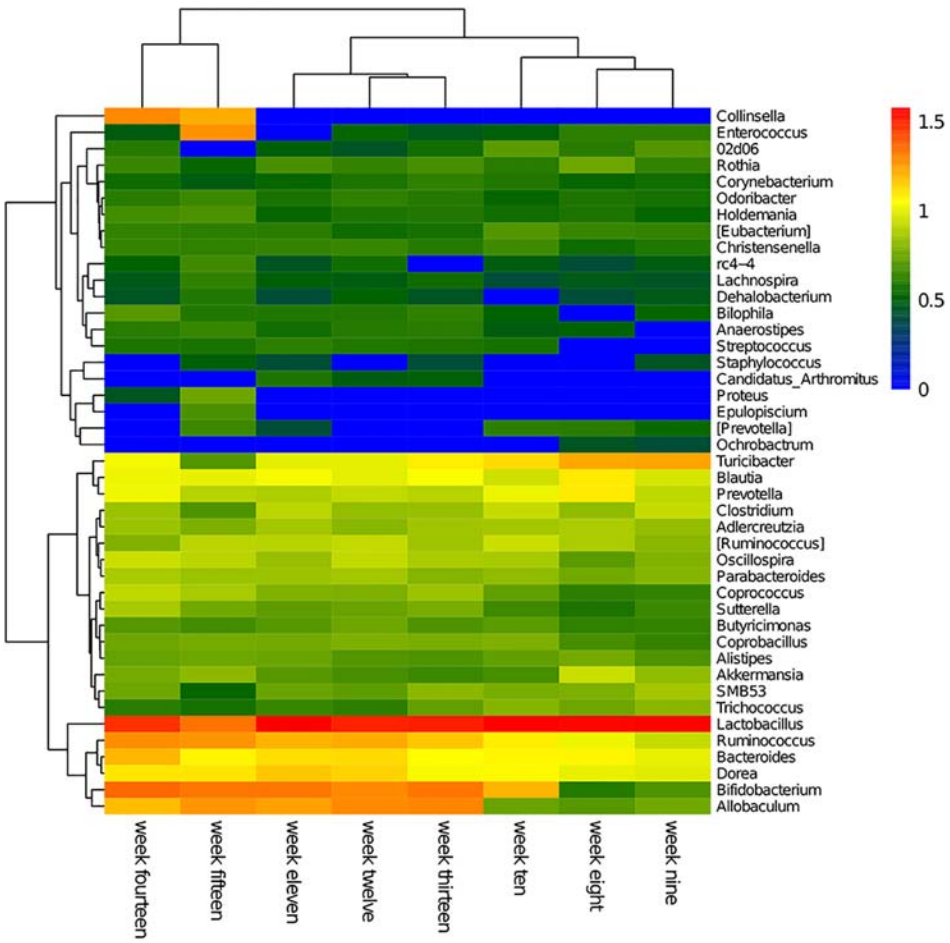


FIGURE 3 | Hierarchical clustering using Euclidean distances was used to construct an inter-week genus-level heatmap. In the figure, red represents high abundance and blue represents low abundance.

(weeks 11–15) grouped into a large cluster (**Figure 7**). To display the number of common and unique OTUs presented in each group during the progression of diabetes, a petal diagram was constructed (**Figure 8A**). It revealed that among all the weeks, ∼306 OTUs were shared. It enabled us to more clearly visualize those OTUs that were distinct for each time scale [ranges from 2 (week 12) to 95 (week 15)] (**Figure 8B**). The dominant phyla of these unique OTUs were *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria*.

Microbial Diversity Initially Increases and Then Decreases With Age and Disease Progression in ZDF Rats

Employing the test for equal proportions (using Pearson's chi-square test statistic), a total of 16 dominant genera ($p < 0.05$) were found in the feces of ZDF rats among the developmental weeks (**Figure 9**). From this relative abundance OTU plot, it is clear that *Lactobacillus* was the predominant genus at 8 weeks of age, along with the presence of *Turicibacter*, *Adlercreutzia*, *Ruminococcus*, *Bacteroides*, *Coprococcus*, *Prevotella*, *Blautia*, *Allobaculum*, *Oscillospira*, *Dorea*, *Clostridium*, *Bifidobacterium*, *Rothia*, *Akkermansia*, and *Trichococcus*. At 9 weeks of age, *Lactobacillus* was also the dominant genus, and abundance of *Turicibacter* was slightly reduced. At 10 weeks of age, *Lactobacillus* continued to increase, *Turicibacter* decreased, but *Bifidobacterium* was significantly present. At 11 weeks of age, *Lactobacillus* and *Bifidobacterium* became the most abundant genera, *Ruminococcus*, *Dorea*, and *Allobaculum* were significantly present, and *Turicibacter* was greatly reduced. The abundance of *Allobaculum* increased from weeks 11 to 15. At week 12, *Lactobacillus*, *Bifidobacterium*, and *Allobaculum* remained the dominant genera until week 13. At week 14, *Lactobacillus*, *Bifidobacterium*, and *Ruminococcus* were the most abundant genera, and *Bacteroides* abundance was significantly elevated. At week 15, *Bifidobacterium* abundance was significantly elevated and it remained the dominant genus along with *Lactobacillus*, *Ruminococcus*, and *Allobaculum*. In brief, *Lactobacillus* was the most abundant genus in feces during the progression of diabetes in ZDF rats. The abundance of *Turicibacter* decreased from weeks 8 to 15. The abundance of *Allobaculum* increased from weeks 11 to 15. We provide the bar plot for the average

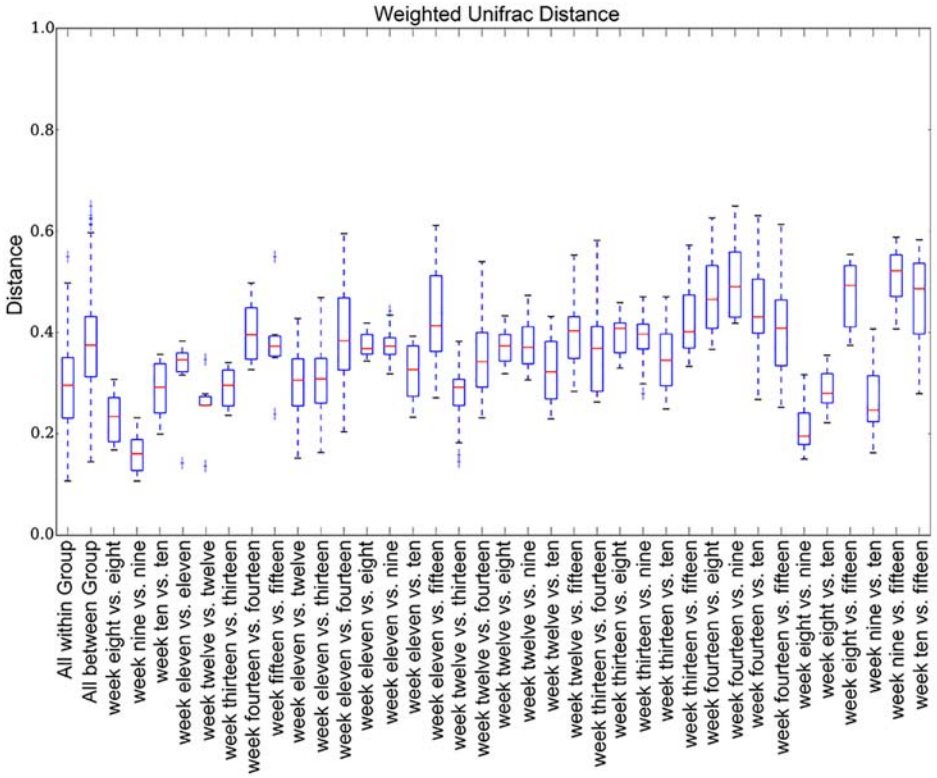


FIGURE 4 | Weighted UniFrac distance box plots. The inter-week weighted UniFrac distances are longer than the intra-week distances.

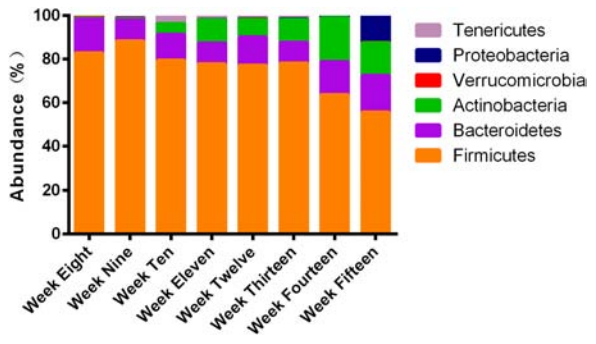


FIGURE 5 | Four phyla: *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria* dominated the majority of ~315 OTUs. 16S rRNA gene sequences were used to establish the identities. Different proportions of phyla can be seen at different stages of diabetes development in ZDF rats. More than 90% of the reads belonged to these four phyla.

abundance of weekly OTUs to clearly visualize the remaining OTUs in **Supplementary Figure S3**. It can be seen that 16 genera accounted for 50–60% of the total genera present per group. The remaining percentage is occupied by low abundance taxa ($n = 62$).

To analyze the genera with the greatest temporal variation, the relative abundances of species at the genus level were employed. This resulted in the selection of 15 genera based on significant differences ($P < 0.05$) (**Figure 10**). These data depicting changes in the abundance levels over time indicate that microbial populations changed significantly over time. Although we could analyze the genera that showed large fluctuations in their abundance levels across the developmental weeks, it must be noted that the genera *Bilophila*, *Proteus*, *Rothia*, and *Streptococcus* had significantly low/negligible abundance levels. These fluctuations with low abundance levels might be attributed to sequencing and/or normalization adjustments. The temporal fluctuations of different microbial communities generally indicate that microbial populations are dynamic during the progression of diabetes over time. We speculate that diet, geography, and other environmental factors play an important role in the development of diabetic microbial communities. Finally, the maximum richness in microbial diversity was obtained in rats at 12 weeks of age.

Physiological Characteristics in ZDF Rats Are Associated With Dysregulated Microbial Taxa

The *Firmicutes*/*Bacteroidetes* (*F*/*B*) ratio is widely used to indicate microbial dysbiosis. Spearman's correlation analysis showed a significant, negative correlation between F/B and age [$R = -0.35$, $P = 0.04$] (**Figure 11A**); however, no significant correlations between F/B and body weight, RBG, food intake, water intake, and rectal temperature were found (**Figures 11B–F**). RBG was strongly and negatively associated with the relative abundance

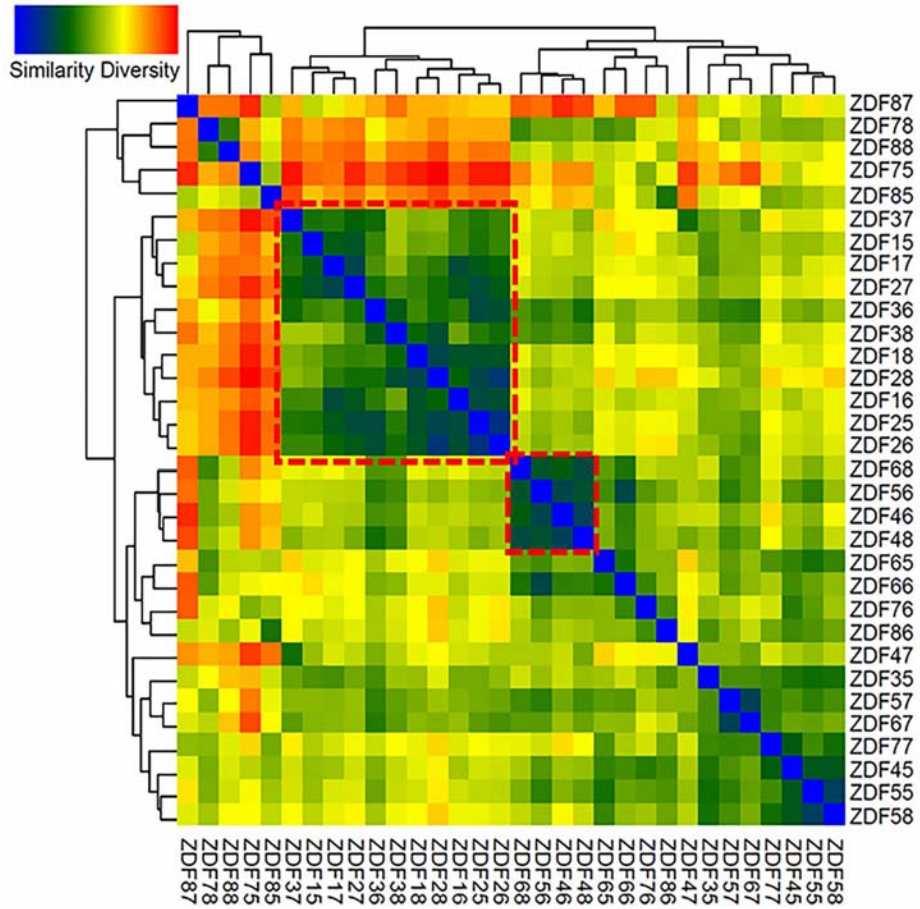


FIGURE 6 | Heatmap generated of the 32 samples. Clustering similarity based on Weighted UniFrac distance matrix data was performed to construct the heatmap; blue in the graph represents high similarity and red represents low similarity. ZDF15–18 represents fecal samples from 8-week-old ZDF rats, ZDF25–28 represents fecal samples from 9-week-old ZDF rats, ZDF35–38 represents fecal samples from 10-week-old ZDF rats, ZDF45–48 represents fecal samples from 11-week-old ZDF rats, ZDF55–58 represents fecal samples from 12-week-old ZDF rats, ZDF65–68 represents fecal samples from 13-week-old ZDF rats, ZDF75–78 represents fecal samples from 14-week-old ZDF rats, and ZDF85–88 represents fecal samples from 15-week old ZDF rats.

values of *Lactobacillus* [$R = -0.42$, $P = 0.02$] and *Turicibacter* [$R = -0.48$, $P = 0.004$] (**Figures 12A,B**) but positively associated with the relative abundance values of *Ruminococcus* [$R = 0.45$, $P = 0.009$] and *Allobaculum* [$R = 0.37$, $P = 0.03$] (**Figures 12C,D**) and SDI [$R = 0.44$, $P = 0.01$] (**Figure 12E**). We found that there was no significant correlation between RBG and relative abundance values of *Bacteroides* [$R = 0.31$, $P = 0.08$], *Akkermansia* [$R = -0.20$, $P = 0.27$], and *Bifidobacterium* [$R = 0.21$, $P = 0.24$] (**Figures 12F–H**). The implications of these associations are unclear and would require further experimentation to demonstrate causality.

DISCUSSION

Microbes play a crucial role in many metabolic-related diseases such as T2DM (Qin et al., 2012). However, systematic studies on the dynamic correlation between microbes and the progression of T2DM are lacking. Therefore, we generated a temporal map of microbial diversity during the progression of T2DM by analyzing the composition of microbes residing in rat feces at different ages. To this end, high-throughput 16S rRNA pyrosequencing was used to study the progressing T2DM fecal microbiome. We used rats of different ages, ranging from 8 to 15 weeks (diabetic stage). We observed that physiological characteristics in ZDF rats, including body weight, food intake, water intake, and RBG increased over time; however, glucose tolerance was impaired and diabetic pathological conditions were aggregated. The phyla *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria* dominated the fecal microbiome during the progression of T2DM. We also demonstrated that *Lactobacillus* and *Turicibacter* are the dominant genera at 8–10 weeks of age, while significant richness and diversity were achieved at 11–12 weeks of age. The maximum diversity was achieved at 12 weeks of age. We believe that these findings significantly improve our understanding of the fecal microbiome during the progression of T2DM.

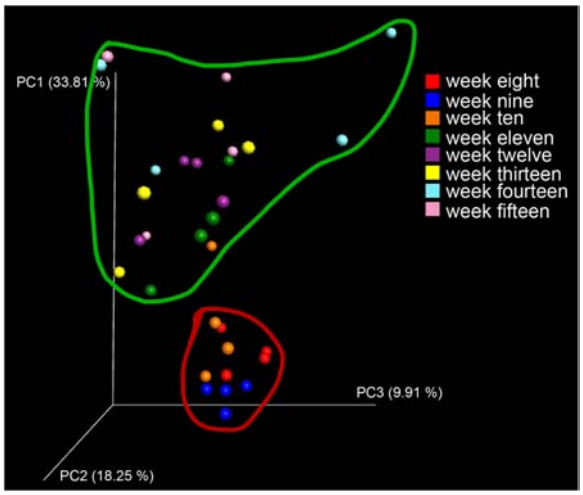
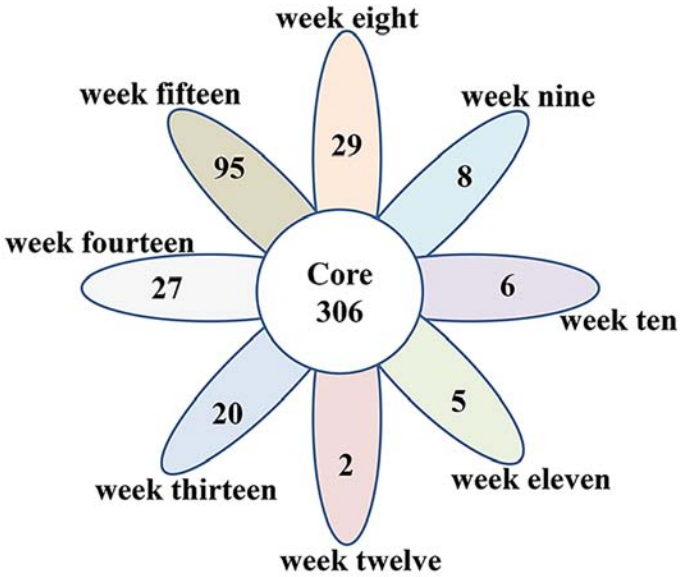


FIGURE 7 | Diversity and distribution of OTUs at different stages of diabetes development. As a measure of beta diversity, PCoA of weighted UniFrac distances (between samples diversity): samples from week 8 (red dots), week 9 (blue dots), and week 10 (brown dots) grouped together into a cluster (when viewed in 3D along the PC3 axis).

Advances in high-throughput sequencing have made it possible to analyze temporal variations in microbial communities based on time series and longitudinal studies. Unique ecological observations relating to the dynamics, stability, and diversity of microbial populations are revealed in these studies. At present, research on temporal data is still rare, and published studies have often focused on only a few time points in many subjects (Horie et al., 2017; Liu et al., 2017). Complex interactions among microbiota may either occur between microorganisms and their niche environment or between microbes. These factors may contribute to the temporal dynamics of microbial populations. In this study, we used bioinformatic strategies to characterize the specific aspects in fecal samples from T2DM. We traced dynamic changes in the rat fecal microbiome during the progression of T2DM by using well-established statistical methods, such as hierarchical clustering and PCoA.

One of the most important findings from this research is that microbial diversity in the rats increased gradually from 8 to 12 weeks of age and slightly decreased from 13 to 15 weeks of age with the progression of T2DM. The diversity at various periods of T2DM was measured by sophisticated indices. Thus, in future studies, we will address whether microbial diversity affects the severity or incidence of diabetes. Another important finding of this study was that, based on weighted UniFrac distance, the fecal microbiome from rats of similar ages were grouped together in the cluster analyses. At all ages, four phyla, namely *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria* dominated the fecal microbiome. Importantly, *Firmicutes* and *Bacteroidetes* were the predominant phyla at all ages. This is of note as about 95% of the human intestinal microbial metabolic profile belongs to *Firmicutes* and *Bacteroidetes*, followed by *Actinobacteria* and *Proteobacteria* (Dicksved et al., 2007; Jernberg et al., 2007; Naseer et al., 2014), suggesting that the human and rat microbiomes are identical in composition at the phylum level. In addition, at the genus level, we observed that *Lactobacillus* was the dominant genus at 8 weeks of age and remained predominant throughout T2DM development. Several reports have indicated that an increase in the abundance of *Lactobacillus* is essential for the prevalence of obesity (Ley et al., 2006; Turnbaugh et al., 2006; Million et al., 2012a,b). Similarly, reports also illustrate the presence of a greater number of *Lactobacillus* in patients with T2DM and ZDF rats, which contributes to the development of chronic inflammation of diabetes (Zeuthen et al., 2006; Sato et al., 2014; Gu et al., 2016). Moreover, *Lactobacillus* is involved in insulin resistance (Le et al., 2012) and is coincident with bile salt hydrolase enzymatic activity, thereby disturbing lipid and glucose metabolism and contributing to T2DM (Tremaroli and Backhed, 2012). We observed a slight decrease in the abundance of *Turicibacter*, a Gram-positive, strictly anaerobic bacterium (Bosshard et al., 2002), in rats at 9 weeks of age. It has also been reported that *Turicibacter* was associated with intestinal butyric acid (Zhong et al., 2015). Butyric acid is a short-chain fatty acid that stimulates insulin secretion in the pancreas, increases insulin sensitivity, and alters insulin signaling (Gao et al., 2009; De Vadder et al., 2014). It has significant functions such as providing anti-obesity effects, reducing metabolic stress, and inhibiting inflammatory reactions (Li et al., 2013; Valvassori et al., 2014). However, the metabolism of *Turicibacter* and its interaction with the host in the intestine are still unclear. *Bifidobacterium* was significantly present in rats at 10 weeks of age. *Bifidobacterium*, a dominant member of the intestinal microbiota and probiotic strain of the phylum *Actinobacteria*, was increased in non-diabetics than in T2DM patients. It has been reported that endotoxemia negatively correlates with *Bifidobacterium* and positively correlates with improved glucose tolerance, glucose-induced insulin secretion, decreased endotoxemia, and adipose tissue proinflammatory cytokines (Cani et al., 2007b). This is because *Bifidobacterium* improves mucosal barrier function, thereby decreasing endotoxin levels (Griffiths et al., 2004; Wang et al., 2006). At 11 weeks of age, *Lactobacillus* and *Bifidobacterium* became the dominant genera, and *Ruminococcus* and *Allobaculum* were significantly present. *Ruminococcus* has been shown to assist gut epithelial cells to absorb sugars, which could contribute to weight gain in the host. Nobel et al. (2015) reported that *Allobaculum* was an important functional phenotype of metabolic dysbiosis. Additionally, it has been reported that *Allobaculum* is the abundant genus in mice that are particularly fed on low-fat and high-fat diets (Ravussin et al., 2012). At 12 weeks of age, *Lactobacillus*, *Bifidobacterium*, and *Allobaculum* remained the dominant genera until week 13. Likewise, *Lactobacillus* was also the dominant genus in 12-week-old TSOD mice (12-week-old TSOD mice exhibit typical clinical status of diabetes) (Horie et al., 2017). At 14 weeks of age, *Bacteroides* was significantly elevated. *Bacteroides* is a Gram-negative bacterium that contains lipopolysaccharide in its cell wall (Finegold et al., 2010). It is known that a large number of Gram-negative bacteria in the intestine may damage the gut barrier, releasing lipopolysaccharide into the bloodstream and triggering a low degree of chronic inflammation (Cani et al., 2007a). Although *Lactobacillus* and *Turicibacter* were the

A



B

Week	No. of OTUs	No. of unique OTUs
Week Eight	1286	29
Week Nine	1310	8
Week Ten	1508	6
Week Eleven	1513	5
Week Twelve	1313	2
Week Thirteen	1553	20
Week Fourteen	1463	27
Week Fifteen	1123	95

FIGURE 8 | The petal diagram reveals common and unique genera associated with different stages of diabetes development. Different colors represent different modules. **(A)** The petal diagram (nodes) at the center of the petal diagram (~306) is shared by all weeks. **(B)** The total number of OTUs and the number of unique OTUs are shown in the table.

predominant genera in 8- to 10-week-old rats, *Bifidobacterium*, *Lactobacillus, Ruminococcus,* and *Allobaculum* were the most abundant genera in 15-week-old rats. One possible reason is that, at the genus level, *Lactobacillus* predominates throughout the progression of T2DM. *Turicibacter* only predominated in the early stage of diabetes in ZDF rats, while the abundance of *Bifidobacterium, Ruminococcus,* and *Allobaculum* increased with the aggravation of the pathological state of diabetes and elevated blood glucose levels in rats (Gu et al., 2016; Kim et al., 2017). Blood glucose levels may also affect the abundance of the bacteria, however, the causal relationship between them is still unclear. Future research needs to prove the relationship between them. In addition, as the rats continued eating high-fat diets, *Allobaculum* may also gradually increase in abundance. These all indicated that, at the genus level, the fecal microbes in the diabetic stage of ZDF rats changed dynamically. Of interest, when compared with previous weeks, a relatively higher diversity was observed at the genus levels at 12 weeks of age, whereas during 13–15 weeks of age, lower diversity was achieved.

Intriguingly, we have observed a gradual decrease in the abundance of *Akkermansia muciniphila* with the progression of diabetes. *A. muciniphila* is an adherent mucin-degrading bacterium that has been proposed to modulate intestinal health, energy balance, and glucose balance (Everard and Cani, 2013). Recent studies uncovered that *A. muciniphila* decreases in prediabetic patients (Yassour et al., 2016; Allinet al., 2018) and has a negative association with T2DM, implying a protective effect on diabetes (Everard et al., 2013; Chen M. et al., 2018; Mithieux, 2018). *A. muciniphila* is found in the feces of rats, and

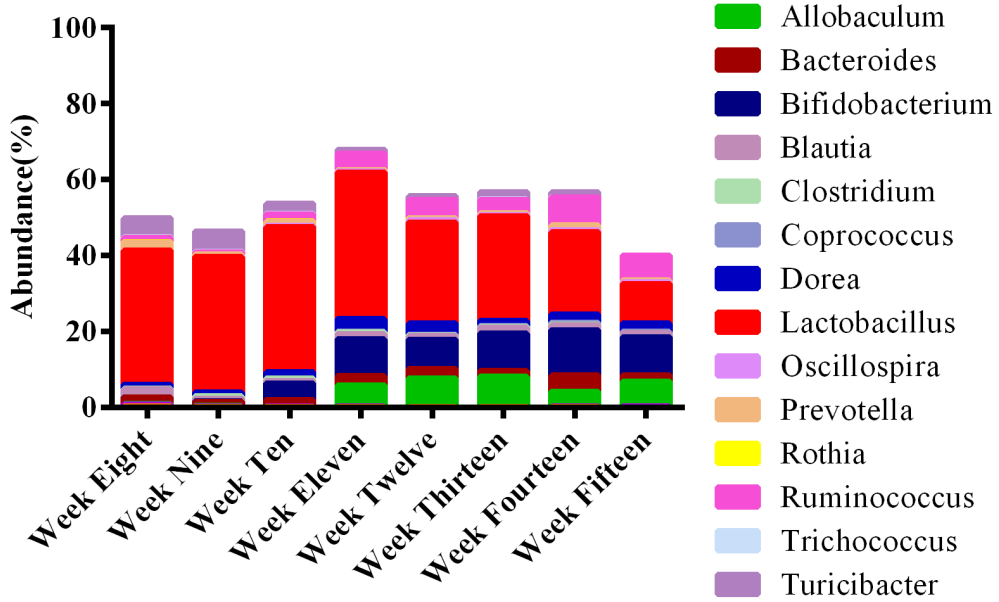


FIGURE 9 | Microbial communities at different stages of diabetes development are dynamic over time. The stacked bar plot reveals that the fecal microbiota was dominated by 16 genera. The average abundance of OTUs per week is represented in bars. *Lactobacillus* was most prominent among the rats at all ages.

a decrease in its abundance is associated with the progression of diabetes. The mechanisms and factors that play an important role in promoting the growth of the bacteria remain unknown and will play an important role in modulation of metabolic diseases.

The presence of *Ruminococcus* in the stool was a surprising finding. Of interest, *Ruminococcus* was earlier detected in the feces and gut flora from adults with T2DM (Everard and Cani, 2013). Since a number of *Ruminococcus* species are known to be associated with metabolic diseases, the identification of *Ruminococcus* to the species level might be critical for further understanding the relationship between *Ruminococcus* and diabetes and its effect on the development of metabolic diseases.

Notably, we found that the *F*/*B* ratio is negatively correlated with age. The *F*/*B* ratio, the ratio of the two largest microbial phyla, has previously been considered to be a sign of obesity and T2DM (Turnbaugh et al., 2009). However, the causality of this transformation of the phyla as an integral part in the health of the organism, and even as a useful biomarker, has recently been questioned (Brown et al., 2012). RBG is negatively associated with *Lactobacillus* and *Turicibacter*, while it is positively correlated with *Ruminococcus*, *Allobaculum* and SDI, suggesting that under conditions of diabetes, *Lactobacillus* and *Turicibacter* may help recover blood glucose levels. There may be mutual influence between the blood glucose levels and microbial diversity; however, precisely how they are affected and whether there is a link in function and causality requires further proof of experimentation. The lack of statistical significance between the *F*/*B* ratio and body weight, food intake, water intake, and RBG, and the lack of statistical significance between RBG levels and *Akkermansia*, *Bacteroides*, and *Bifidobacterium*, may be due to variability among individuals and a small sample size. Although a few studies have shown a correlation between specific physiological characteristics and specific gut microbes, to the best of our knowledge, this is the first study to attempt to correlate the dynamic physiological characteristics of 8–15-week-old ZDF rats with microbes. Further investigations are required with a greater number of animals or a human cohort to verify the results of this study and to determine the possible underlying mechanisms.

T2DM is a complex metabolic disorder. Beyond the widely-accepted concept that genetic factors play an important role in diabetes susceptibility, growing evidence has demonstrated that environmental factors (such as commensal bacteria, chemicals, diet, and viruses) may also modify diabetes development. Of these factors, the gut microbiota has been shown to play an important role in influencing the progression of T2DM. This has been supported by results from both human research and animal studies, especially the discordant incidence of diabetes in monozygotic twins who are genetically identical (Tai et al., 2015). *Lactobacillus* might be used as one of the genera in experiments, showing a role for the microbiome during the progression in T2DM. We also observed that among groups of rats of different ages, *Firmicutes* and *Bacteroides* are the dominant phyla. These observations confirm findings from patients with diabetes, where these phyla were found in the feces of diabetics (Dicksved et al., 2007; Jernberg et al., 2007; Naseer et al., 2014). In this study, we observed that the microbiome of rats was predominated by the genera *Lactobacillus*, *Turicibacter*, *Bifidobacterium*, *Ruminococcus*, *Allobaculum*, and *Bacteroides*. Studies in humans

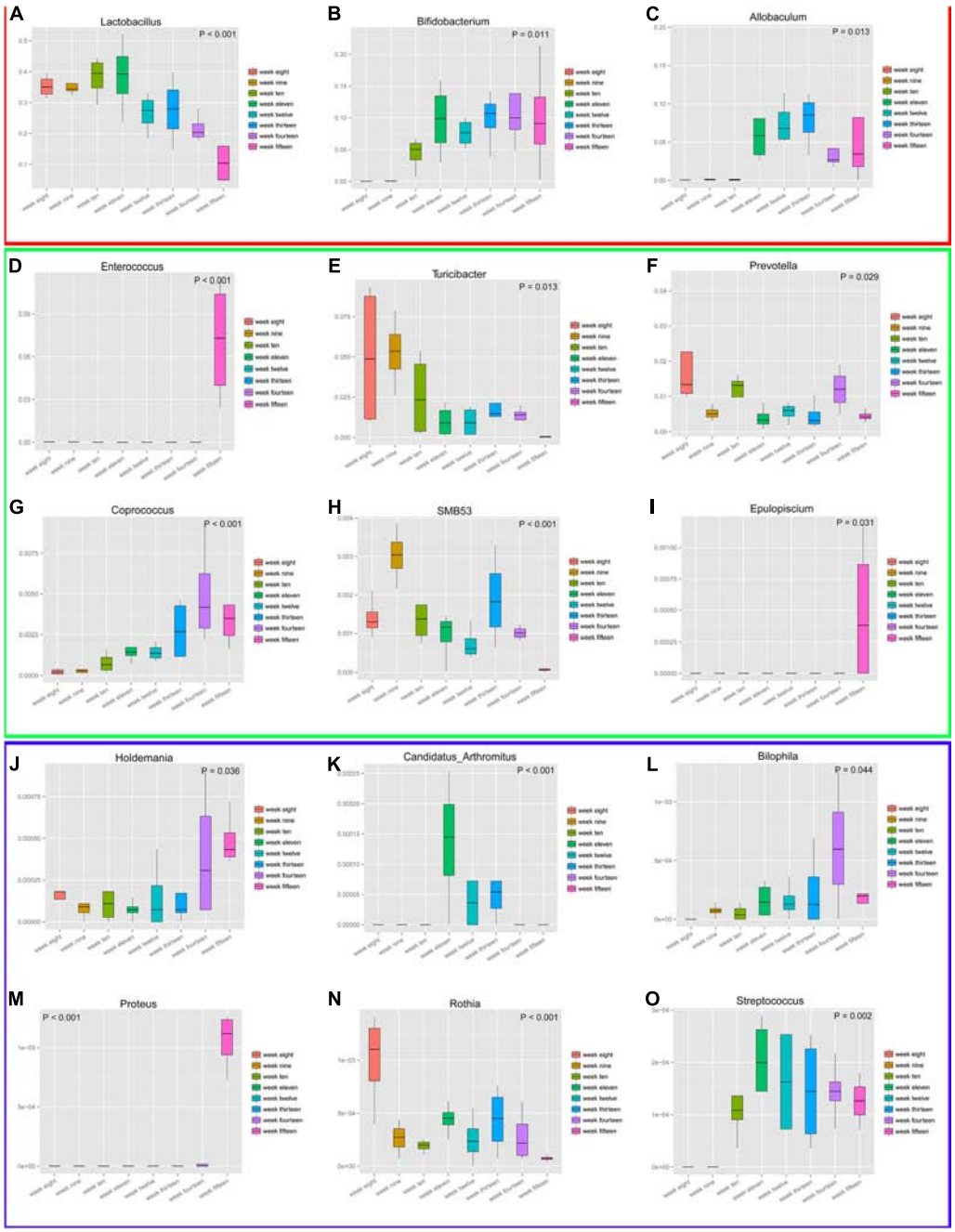


FIGURE 10 | Fifteen genera experienced maximum temporal fluctuations in abundance levels at different stages of diabetes development. These 15 genera **(A–O)** depict highly dynamic variations. *P*-values ($P < 0.5$) are shown on plot corners. Box plots display the following values: the *Y*-axis represents relative abundance of the genus, the *X*-axis represents time grouping; middle box line, represents the median; the upper and lower whiskers represent 1.5 times IQR beyond the upper and lower quartiles, respectively; and dots represent outlier values. Genera *Bilophila*, *Proteus*, *Rothia*, and *Streptococcus* have low abundance levels. For better inspection, these plots have been divided into three parts (red, green, and blue) that reflect their relative abundances.

suggest that the human fecal microbiome is primarily dominated by *Bifidobacterium*, *Bacteroides*, *Escherichia*, *Intestinibacter*, *Prevotella*, *A. muciniphila*, *Blautia*, and *Ruminococcus* (Moreno-Indias et al., 2014; Wu et al., 2017). These data seem to indicate that the rat fecal microbiome has some similarities with the human fecal microbiome while harboring some other genera. This observation indicates that the ZDF rat can be used as a model for studying the T2DM microbiome. Analyzing the changes in the rat fecal microbiome and comparing them with the available data from human clinical studies will be interesting. In addition, increasing evidence indicates miRNAs have close associations with diabetes, so miRNA biomarkers

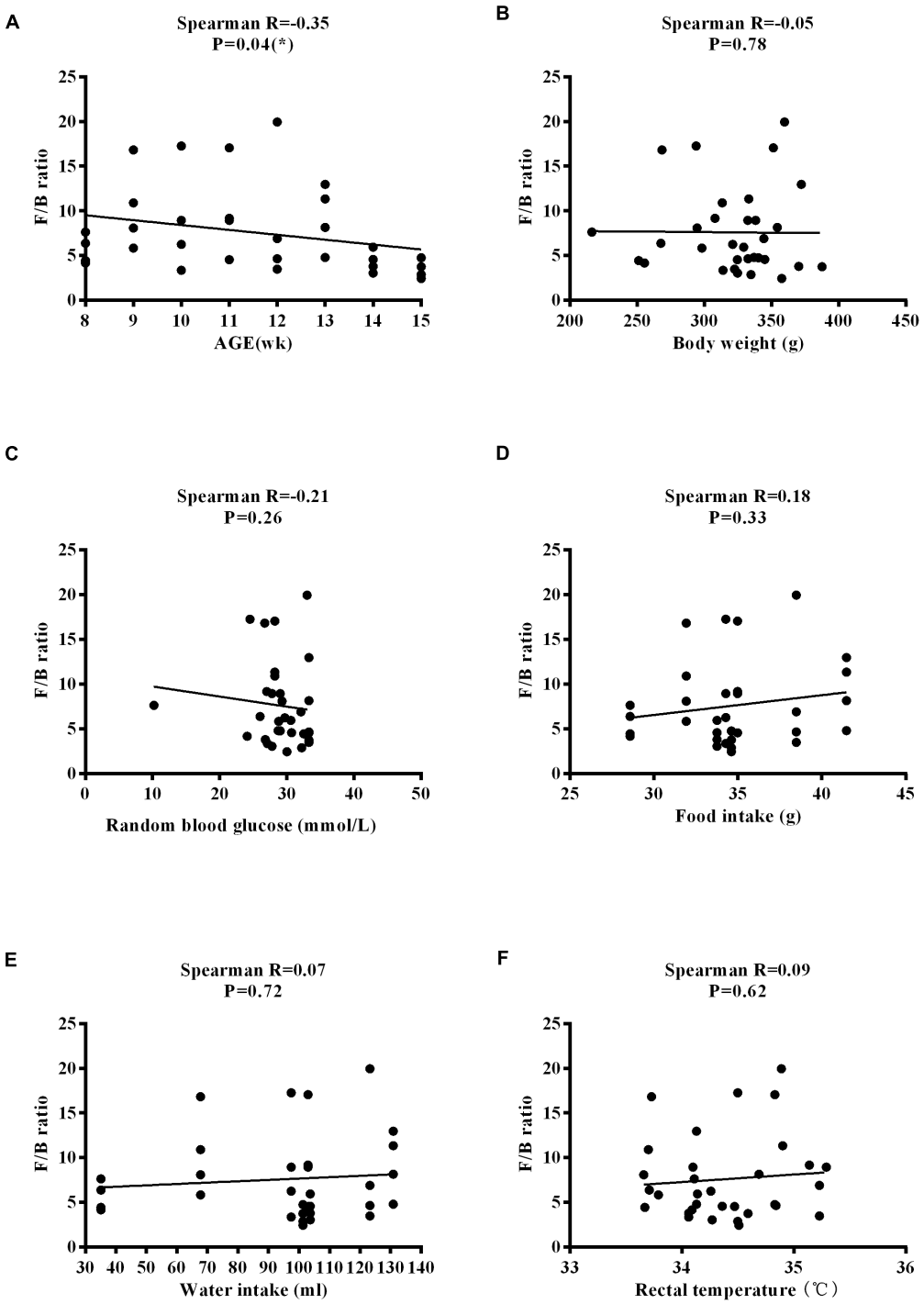


FIGURE 11 | Correlation between the *F/B* ratio and physiological characteristics in ZDF rats. **(A)** Age. **(B)** Body weight. **(C)** Random blood glucose. **(D)** Food intake. **(E)** Water intake. **(F)** Rectal temperature.

will be particularly useful in early diagnostics of diabetes (Chen et al., 2017; Chen X. et al., 2018; Hu et al., 2018; Zhao et al., 2018a,b). It would be meaningful to correlate miRNA biomarkers with gut microbes in T2DM, but this is still beyond the scope of this study.

Although our research has monitored the changes in microbiome composition and the diversity of feces in ZDF rats with age and disease progression, important questions remain unanswered. These include whether the fecal microbiome is influenced by sex or diet. A number of earlier studies have

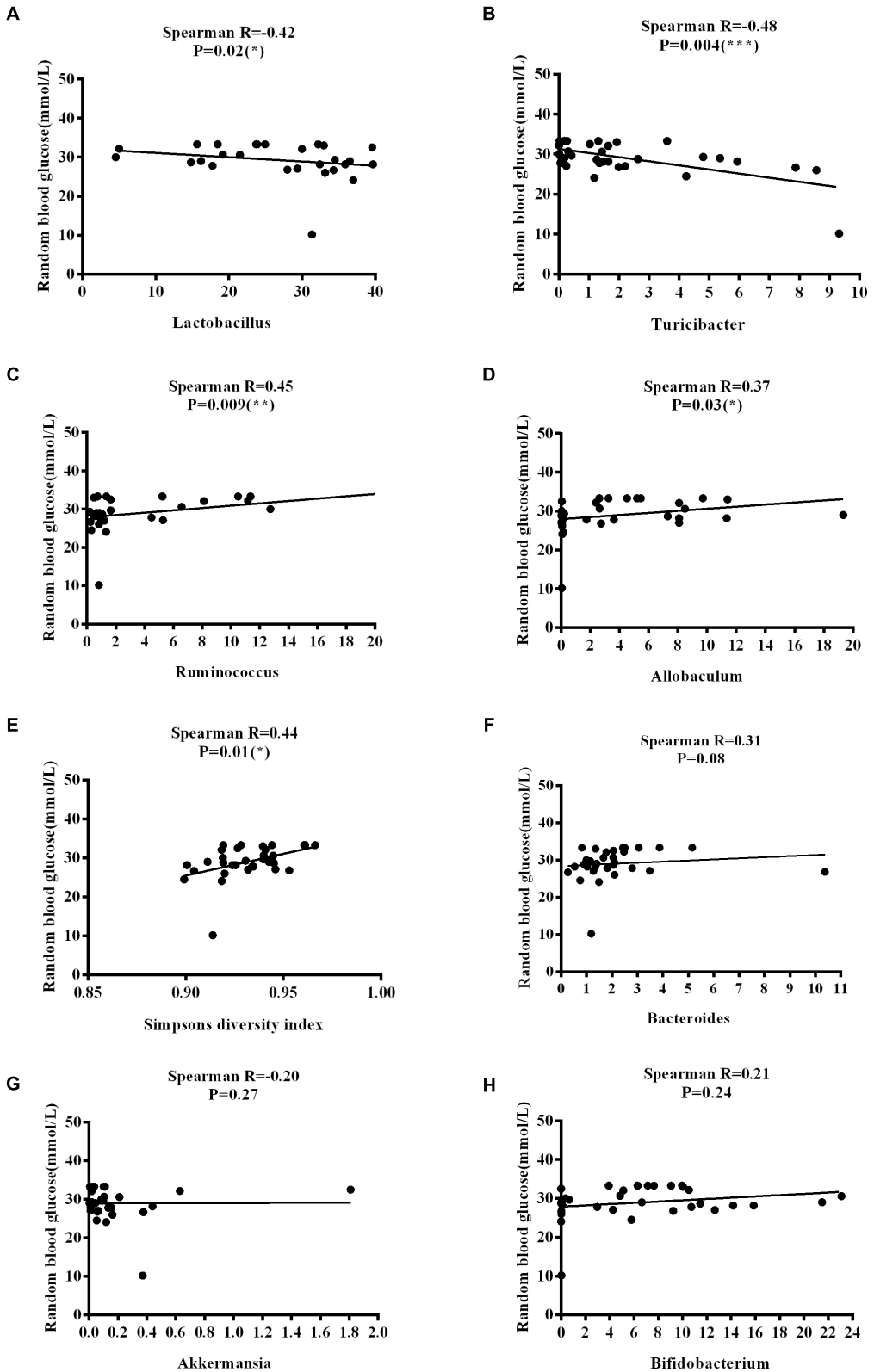


FIGURE 12 | Correlations between random blood glucose levels and variation in microbial communities and Simpson's diversity index. **(A)** *Lactobacillus*. **(B)** *Turicibacter*. **(C)** *Ruminococcus*. **(D)** *Allobaculum*. **(E)** Simpson's diversity index. **(F)** *Bacteroides*. **(G)** *Akkermansia*. **(H)** *Bifidobacterium*.

shown that sex may influence the fecal microbiome (Shastri et al., 2015; Fields et al., 2018). Diet can also affect the intestinal flora, especially high-fat diets (Daniel et al., 2014; Chi et al., 2018; He et al., 2018) Thus, one of the weaknesses of the present study is that we did not take into account the effects of sex or diet on the rat fecal microbiome. Another limitation of this study is that our results are based on a small sample size. Further validation in a larger number of animals or a human cohort is needed. Moreover, we have not demonstrated a causal relationship between microbiota and diabetes. Future work to establish causality would involve the isolation of specific taxa and transfer of anaerobically cultured clones into germ-free animals to demonstrate the development of diabetes in recipient animals. Unfortunately, there was no negative control group included to dissect the specific correlation between microbial changes and disease progression. Future studies should include negative control studies to better understand the correlation between microbial changes and disease progression. The main strength of this research is that fecal microbiome composition was associated with age and the progression of diabetes. Toward this, fecal samples in the rats of different ages were collected and fecal microbiome analysis was performed. To conclude the association between the diabetic microbiome and age and disease progression in rats, we performed rigorous analyses.

This research differs from other research projects which address the effect of the gut microbes on diabetes. In this study, we monitored the changes in the fecal microbiome with the growth and disease progression of T2DM. However, many other confounding factors may affect fecal microbes, including stress and feed type, as in the gut microbiome (Hufeldt et al., 2010). In summary, we have monitored the changes in the fecal microbiome in ZDF rats from 8 to 15 weeks of age by using deep sequencing. This analysis suggests that the microbial composition is associated with the age and progression of diabetes in rats.

CONCLUSION

Other research has implicated the microbiome in playing an important role in metabolic diseases such as diabetes. However, there is a lack of time-resolved microbial changes during the progression of diabetes. This understanding is crucial for creating new interventions for curing metabolic diseases such as diabetes. In this study, we monitored changes in the fecal microbiome during the progression of diabetes from 8 to 15 weeks of age. The fecal microbiome in rats was highly dynamic and underwent major changes during the progression of diabetes. The determined time-dependent alteration of the fecal microbiome supports further investigation to determine whether *Lactobacillus*, *Turicibacter*, *Bifidobacterium*, *Allobaculum*, *Ruminococcus*, and *Akkermansia* may play functional roles in the progression of diabetes before any intervention can be considered.

AUTHOR CONTRIBUTIONS

LBZ and XL conceived the idea, directed the project, and designed the experiments. HX, WZ, and LJZ performed the experiments, obtained the samples, and acquired the data. WZ analyzed the data and wrote the manuscript. LBZ, XL, and WZ edited the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by The Key Project of the National Natural Science Foundation of China (Nos. 81730111 and 81230084), the Traditional Chinese Medicine Leading Intelligence Project of Jiangsu Province (No. SLJ0227), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX17_1314).

ACKNOWLEDGMENTS

We would like to thank Shanghai Personal Biotechnology Co., Ltd. (Shanghai, China) for providing sequencing services and helpful discussions pertaining to the sequencing and data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00232/full#supplementary-material

REFERENCES

Allin, K. H., Tremaroli, V., Caesar, R., Jensen, B. A. H., Damgaard, M. T. F., Bahl, M. I., et al. (2018). Aberrant intestinal microbiota in individuals with prediabetes. *Diabetologia* 61, 810–820. doi: 10.1007/s00125-018-4550-1

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Bagarolli, R. A., Tobar, N., Oliveira, A. G., Araujo, T. G., Carvalho, B. M., Rocha, G. Z., et al. (2017). Probiotics modulate gut microbiota and improve insulin sensitivity in DIO mice. *J. Nutr. Biochem.* 50, 16–25. doi: 10.1016/j.jnutbio.2017.08.006

Bindels, L. B., Segura Munoz, R. R., Gomes-Neto, J. C., Mutemberezi, V., Martinez, I., Salazar, N., et al. (2017). Resistant starch can improve insulin sensitivity independently of the gut microbiota. *Microbiome* 5:12. doi: 10.1186/s40168-017-0230-5

Bosshard, P. P., Zbinden, R., and Altwegg, M. (2002). *Turicibacter sanguinis* gen. nov., sp. nov., a novel anaerobic, Gram-positive bacterium. *Int. J. Syst. Evol. Microbiol.* 52, 1263–1266.

Brown, K., Decoffe, D., Molcan, E., and Gibson, D. L. (2012). Diet-induced dysbiosis of the intestinal microbiota and the effects on immunity and disease. *Nutrients* 4, 1095–1119. doi: 10.3390/nu4081095

Cani, P. D., Amar, J., Iglesias, M. A., Poggi, M., Knauf, C., Bastelica, D., et al. (2007a). Metabolic endotoxemia initiates obesity and insulin resistance. *Diabetes* 56, 1761–1772. doi: 10.2337/db06-1491

Cani, P. D., Neyrinck, A. M., Fava, F., Knauf, C., Burcelin, R. G., Tuohy, K. M., et al. (2007b). Selective increases of bifidobacteria in gut microflora improve

high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia. *Diabetologia* 50, 2374–2383.

Caparros-Martin, J. A., Lareu, R. R., Ramsay, J. P., Peplies, J., Reen, F. J., Headlam, H. A., et al. (2017). Statin therapy causes gut dysbiosis in mice through a PXR-dependent mechanism. *Microbiome* 5:95. doi: 10.1186/s40168-017-0312-4

Chen, H., and Jiang, W. (2014). Application of high-throughput sequencing in understanding human oral microbiome related with health and disease. *Front. Microbiol.* 5:508. doi: 10.3389/fmicb.2014.00508

Chen, M., Liao, Z., Lu, B., Wang, M., Lin, L., Zhang, S., et al. (2018). Huang-Lian-Jie-Du-Decoction ameliorates hyperglycemia and insulin resistant in association with gut microbiota modulation. *Front. Microbiol.* 9:2380. doi: 10.3389/fmicb.2018.02380

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018). BNPMDA: bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2017). MicroRNAs and complex diseases: from experimental results to computational models. *Brief Bioinform.* [Epub ahead of print].

Chi, Y., Lin, Y., Zhu, H., Huang, Q., Ye, G., and Dong, S. (2018). PCBs-high-fat diet interactions as mediators of gut microbiota dysbiosis and abdominal fat accumulation in female mice. *Environ. Pollut.* 239, 332–341. doi: 10.1016/j.envpol.2018.04.001

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., et al. (2018). IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* 138, 271–281. doi: 10.1016/j.diabres.2018.02.023

Cox, L. M., Yamanishi, S., Sohn, J., Alekseyenko, A. V., Leung, J. M., Cho, I., et al. (2014). Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell* 158, 705–721. doi: 10.1016/j.cell.2014.05.052

Da Silva, B. A., Bjorbaek, C., Uotani, S., and Flier, J. S. (1998). Functional properties of leptin receptor isoforms containing the gln– > pro extracellular domain mutation of the fatty rat. *Endocrinology* 139, 3681–3690. doi: 10.1210/endo.139.9.6168

Daniel, H., Gholami, A. M., and Berry, D. (2014). High-fat diet alters gut microbiota physiology in mice. *ISME J.* 8, 295–308. doi: 10.1038/ismej.2013.155

De Lemos, E. T., Reis, F., Baptista, S., Pinto, R., Sepodes, B., Vala, H., et al. (2007). Exercise training is associated with improved levels of C-reactive protein and adiponectin in ZDF (type 2) diabetic rats. *Med. Sci. Monit.* 13, Br168–Br174.

De Vadder, F., Kovatcheva-Datchary, P., Goncalves, D., Vinera, J., Zitoun, C., Duchampt, A., et al. (2014). Microbiota-generated metabolites promote metabolic benefits via gut-brain neural circuits. *Cell* 156, 84–96. doi: 10.1016/j.cell.2013.12.016

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

Dicksved, J., Floistrup, H., Bergstrom, A., Rosenquist, M., Pershagen, G., Scheynius, A., et al. (2007). Molecular fingerprinting of the fecal microbiota of children raised according to different lifestyles. *Appl. Environ. Microbiol.* 73, 2284–2289. doi: 10.1128/AEM.02223-06

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

Everard, A., Belzer, C., Geurts, L., Ouwerkerk, J. P., Druart, C., Bindels, L. B., et al. (2013). Cross-talk between *Akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 9066–9071. doi: 10.1073/pnas.1219451110

Everard, A., and Cani, P. D. (2013). Diabetes, obesity and gut microbiota. *Best Pract. Res. Clin. Gastroenterol.* 27, 73–83. doi: 10.1016/j.bpg.2013.03.007

Fields, C. T., Chassaing, B., Paul, M. J., Gewirtz, A. T., and De Vries, G. J. (2018). Vasopressin deletion is associated with sex-specific shifts in the gut microbiome. *Gut Microbes* 9, 13–25. doi: 10.1080/19490976.2017.1356557

Finegold, S. M., Dowd, S. E., Gontcharova, V., Liu, C., Henley, K. E., Wolcott, R. D., et al. (2010). Pyrosequencing study of fecal microflora of autistic and control children. *Anaerobe* 16, 444–453. doi: 10.1016/j.anaerobe.2010.06.008

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., et al. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 528, 262–266. doi: 10.1038/nature15766

Gao, Z., Yin, J., Zhang, J., Ward, R. E., Martin, R. J., Lefevre, M., et al. (2009). Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes* 58, 1509–1517. doi: 10.2337/db08-1637

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi: 10.1126/science.1124234

Goldsmith, F., Guice, J., Page, R., Welsh, D. A., Taylor, C. M., Blanchard, E. E., et al. (2017). Obese ZDF rats fermented resistant starch with effects on gut microbiota but no reduction in abdominal fat. *Mol. Nutr. Food Res.* 61:1501025. doi: 10.1002/mnfr.201501025

Griffiths, E. A., Duffy, L. C., Schanbacher, F. L., Qiao, H., Dryja, D., Leavens, A., et al. (2004). In vivo effects of bifidobacteria and lactoferrin on gut endotoxin concentration and mucosal immunity in Balb/c mice. *Dig. Dis. Sci.* 49, 579–589. doi: 10.1023/B:DDAS.0000026302.92898.ae

Gu, C., Yang, Y., Xiang, H., Li, S., Liang, L., Sui, H., et al. (2016). Deciphering bacterial community changes in zucker diabetic fatty rats based on 16S rRNA gene sequences analysis. *Oncotarget* 7, 48941–48952. doi: 10.18632/oncotarget.10597

Gu, C., Zhou, W., Wang, W., Xiang, H., Xu, H., Liang, L., et al. (2017). ZiBuPiYin recipe improves cognitive decline by regulating gut microbiota in Zucker diabetic fatty rats. *Oncotarget* 8, 27693–27703. doi: 10.18632/oncotarget.14611

He, C., Cheng, D., Peng, C., Li, Y., Zhu, Y., and Lu, N. (2018). High-fat diet induces dysbiosis of gastric microbiota prior to gut microbiota in association with metabolic disorders in mice. *Front. Microbiol.* 9:639. doi: 10.3389/fmicb.2018.00639

Horie, M., Miura, T., Hirakata, S., Hosoyama, A., Sugino, S., Umeno, A., et al. (2017). Comparative analysis of the intestinal flora in type 2 diabetes and nondiabetic mice. *Exp. Anim.* 66, 405–416. doi: 10.1538/expanim.17-0021

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., and Zhao, Q. (2018). HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Hufeldt, M. R., Nielsen, D. S., Vogensen, F. K., Midtvedt, T., and Hansen, A. K. (2010). Variation in the gut microbiota of laboratory mice is related to both genetic and environmental factors. *Comp. Med.* 60, 336–347.

Jernberg, C., Lofmark, S., Edlund, C., and Jansson, J. K. (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *ISME J.* 1, 56–66. doi: 10.1038/ismej.2007.3

Jia, L., Li, D., Feng, N., and Shamoon, M. (2017). Anti-diabetic effects of *Clostridium butyricum* CGMCC0313.1 through promoting the growth of gut butyrate-producing bacteria in type 2 diabetic mice. *Sci. Rep.* 7:7046. doi: 10.1038/s41598-017-07335-0

Jourdan, T., Godlewski, G., Cinar, R., Bertola, A., Szanda, G., Liu, J., et al. (2013). Activation of the Nlrp3 inflammasome in infiltrating macrophages by endocannabinoids mediates beta cell loss in type 2 diabetes. *Nat. Med.* 19, 1132–1140. doi: 10.1038/nm.3265

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergstrom, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi: 10.1038/nature12198

Kim, T., Holleman, C. L., Ptacek, T., Morrow, C. D., and Habegger, K. M. (2017). Duodenal endoluminal barrier sleeve alters gut microbiota of ZDF rats. *Int. J. Obes.* 41, 381–389. doi: 10.1038/ijo.2016.224

Le, K. A., Li, Y., Xu, X., Yang, W., Liu, T., Zhao, X., et al. (2012). Alterations in fecal *Lactobacillus* and *Bifidobacterium* species in type 2 diabetic patients in Southern China population. *Front. Physiol.* 3:496. doi: 10.3389/fphys.2012.00496

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. P., Chen, X., and Li, M. Q. (2013). Butyrate alleviates metabolic impairments and protects pancreatic beta cell function in pregnant mice with obesity. *Int. J. Clin. Exp. Pathol.* 6, 1574–1584.

Liu, F., Ling, Z., Xiao, Y., Lv, L., Yang, Q., Wang, B., et al. (2017). Dysbiosis of urinary microbiota is positively correlated with type 2 diabetes mellitus. *Oncotarget* 8, 3798–3810. doi: 10.18632/oncotarget.14028

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

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

Mengual, L., Roura, P., Serra, M., Montasell, M., Prieto, G., and Bonet, S. (2010). Multifactorial control and treatment intensity of type-2 diabetes in primary care settings in Catalonia. *Cardiovasc. Diabetol.* 9:14. doi: 10.1186/1475-2840-9-14

Million, M., Angelakis, E., Paul, M., Armougom, F., Leibovici, L., and Raoult, D. (2012a). Comparative meta-analysis of the effect of *Lactobacillus* species on weight gain in humans and animals. *Microb. Pathog.* 53, 100–108. doi: 10.1016/j.micpath.2012.05.007

Million, M., Maraninchi, M., Henry, M., Armougom, F., Richet, H., Carrieri, P., et al. (2012b). Obesity-associated gut microbiota is enriched in Lactobacillus reuteri and depleted in *Bifidobacterium animalis* and *Methanobrevibacter smithii*. *Int. J. Obes.* 36, 817–825. doi: 10.1038/ijo.2011.153

Mithieux, G. (2018). Does *Akkermansia muciniphila* play a role in type 1 diabetes? *Gut* 67, 1373–1374. doi: 10.1136/gutjnl-2017-315732

Moreno-Indias, I., Cardona, F., Tinahones, F. J., and Queipo-Ortuno, M. I. (2014). Impact of the gut microbiota on the development of obesity and type 2 diabetes mellitus. *Front. Microbiol.* 5:190. doi: 10.3389/fmicb.2014.00190

Naseer, M. I., Bibi, F., Alqahtani, M. H., Chaudhary, A. G., Azhar, E. I., Kamal, M. A., et al. (2014). Role of gut microbiota in obesity, type 2 diabetes and Alzheimer's disease. *CNS Neurol. Disord. Drug Targets* 13, 305–311. doi: 10.2174/18715273113126660147

Nobel, Y. R., Cox, L. M., Kirigin, F. F., Bokulich, N. A., Yamanishi, S., Teitler, I., et al. (2015). Metabolic and metagenomic outcomes from early-life pulsed antibiotic treatment. *Nat. Commun.* 6:7486. doi: 10.1038/ncomms8486

Phillips, M. S., Liu, Q., Hammond, H. A., Dugan, V., Hey, P. J., Caskey, C. J., et al. (1996). Leptin receptor missense mutation in the fatty Zucker rat. *Nat. Genet.* 13, 18–19. doi: 10.1038/ng0596-18

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. doi: 10.1038/nature11450

Ravussin, Y., Koren, O., Spor, A., Leduc, C., Gutman, R., Stombaugh, J., et al. (2012). Responses of gut microbiota to diet composition and weight loss in lean and obese mice. *Obesity* 20, 738–747. doi: 10.1038/oby.2011.111

Sato, J., Kanazawa, A., Ikeda, F., Yoshihara, T., Goto, H., Abe, H., et al. (2014). Gut dysbiosis and detection of "live gut bacteria" in blood of Japanese patients with type 2 diabetes. *Diabetes Care* 37, 2343–2350. doi: 10.2337/dc13-2817

Saxena, R., Elbers, C. C., Guo, Y., Peter, I., Gaunt, T. R., Mega, J. L., et al. (2012). Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. *Am. J. Hum. Genet.* 90, 410–425. doi: 10.1016/j.ajhg.2011.12.022

Shastri, P., Mccarville, J., Kalmokoff, M., Brooks, S. P., and Green-Johnson, J. M. (2015). Sex differences in gut fermentation and immune parameters in rats fed an oligofructose-supplemented diet. *Biol. Sex Differ.* 6:13. doi: 10.1186/s13293-015-0031-0

Szokol, M., Priksz, D., Bombicz, M., Varga, B., Kovacs, A., Fulop, G. A., et al. (2017). Long term osmotic mini pump treatment with alpha-MSH improves myocardial function in zucker diabetic fatty rats. *Molecules* 22:E1702. doi: 10.3390/molecules22101702

Tai, N., Wong, F. S., and Wen, L. (2015). The role of gut microbiota in the development of type 1, type 2 diabetes mellitus and obesity. *Rev. Endocr. Metab. Disord.* 16, 55–65. doi: 10.1007/s11154-015-9309-0

Tremaroli, V., and Backhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature* 489, 242–249. doi: 10.1038/nature11552

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, 480–484. doi: 10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

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

Van Bree, B. W., Lenaers, E., Nabben, M., Briede, J. J., Jorgensen, J. A., Schaart, G., et al. (2016). A genistein-enriched diet neither improves skeletal muscle oxidative capacity nor prevents the transition towards advanced insulin resistance in ZDF rats. *Sci. Rep.* 6:22854. doi: 10.1038/srep22854

Wang, Z., Xiao, G., Yao, Y., Guo, S., Lu, K., and Sheng, Z. (2006). The role of bifidobacteria in gut barrier function after thermal injury in rats. *J. Trauma* 61, 650–657. doi: 10.1097/01.ta.0000196574.70614.27

Wessels, B., Ciapaite, J., Van Den Broek, N. M., Houten, S. M., Nicolay, K., and Prompers, J. J. (2015). Pioglitazone treatment restores in vivo muscle oxidative capacity in a rat model of diabetes. *Diabetes Obes. Metab.* 17, 52–60. doi: 10.1111/dom.12388

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

Wu, H., Esteve, E., Tremaroli, V., Khan, M. T., Caesar, R., Manneras-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, 850–858. doi: 10.1038/nm.4345

Yamashita, T., Murakami, T., Iida, M., Kuwajima, M., and Shima, K. (1997). Leptin receptor of Zucker fatty rat performs reduced signal transduction. *Diabetes* 46, 1077–1080. doi: 10.2337/diab.46.6.1077

Yano, J. M., Yu, K., Donaldson, G. P., Shastri, G. G., Ann, P., Ma, L., et al. (2015). Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell* 161, 264–276. doi: 10.1016/j.cell.2015.02.047

Yassour, M., Lim, M. Y., Yun, H. S., Tickle, T. L., Sung, J., Song, Y. M., et al. (2016). Sub-clinical detection of gut microbial biomarkers of obesity and type 2 diabetes. *Genome Med.* 8:17. doi: 10.1186/s13073-016-0271-6

Yokoi, N., Hoshino, M., Hidaka, S., Yoshida, E., Beppu, M., Hoshikawa, R., et al. (2013). A novel rat model of type 2 diabetes: the zucker fatty diabetes mellitus ZFDM rat. *J. Diabetes Res.* 2013, 103731. doi: 10.1155/2013/103731

Zeuthen, L. H., Christensen, H. R., and Frokiaer, H. (2006). Lactic acid bacteria inducing a weak interleukin-12 and tumor necrosis factor alpha response in human dendritic cells inhibit strongly stimulating lactic acid bacteria but act synergistically with gram-negative bacteria. *Clin. Vaccine Immunol.* 13, 365–375. doi: 10.1128/CVI.13.3.365-375.2006

Zhao, Q., Yu, H., Ming, Z., Hu, H., Ren, G., and Liu, H. (2018a). The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. *Mol. Ther. Nucleic Acids* 13, 464–471. doi: 10.1016/j.omtn.2018.09.020

Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018b). IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zhong, Y., Nyman, M., and Fak, F. (2015). Modulation of gut microbiota in rats fed high-fat diets by processing whole-grain barley to barley malt. *Mol. Nutr. Food Res.* 59, 2066–2076. doi: 10.1002/mnfr.201500187

Conflict of Interest Statement: 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.

Copyright © 2019 Zhou, Xu, Zhan, Lu and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

TECHNOLOGY REPORT
published: 22 February 2019
doi: 10.3389/fmicb.2019.00288

Web-gLV: A Web Based Platform for Lotka-Volterra Based Modeling and Simulation of Microbial Populations

Bhusan K. Kuntal[1,2,3], *Chetan Gadgil*[2,3,4] *and Sharmila S. Mande*[1]*

[1] *Bio-Sciences R&D Division, TCS Research, Tata Consultancy Services Ltd., Pune, India,* [2] *CSIR-National Chemical Laboratory, Chemical Engineering and Process Development Division, Pune, India,* [3] *CSIR-National Chemical Laboratory, Academy of Scientific and Innovative Research, Pune, India,* [4] *CSIR-Institute of Genomics and Integrative Biology, New Delhi, India*

The affordability of high throughput DNA sequencing has allowed us to explore the dynamics of microbial populations in various ecosystems. Mathematical modeling and simulation of such microbiome time series data can help in getting better understanding of bacterial communities. In this paper, we present Web-gLV—a GUI based interactive platform for generalized Lotka-Volterra (gLV) based modeling and simulation of microbial populations. The tool can be used to generate the mathematical models with automatic estimation of parameters and use them to predict future trajectories using numerical simulations. We also demonstrate the utility of our tool on few publicly available datasets. The case studies demonstrate the ease with which the current tool can be used by biologists to model bacterial populations and simulate their dynamics to get biological insights. We expect Web-gLV to be a valuable contribution in the field of ecological modeling and metagenomic systems biology.

Keywords: microbiome, modeling, numerical-simulation, web-server, time-series, visualization, lotka-volterra, microbial population

OPEN ACCESS

Edited by:
Rachel Susan Poretsky, University of Illinois at Chicago, United States

Reviewed by:
Tom O. Delmont, Genoscope, France
Uwe C. Tauber, Virginia Tech, United States

***Correspondence:**
Sharmila S. Mande
sharmila.mande@tcs.com

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

Received: *03 October 2018*
Accepted: *04 February 2019*
Published: *22 February 2019*

Citation:
Kuntal BK, Gadgil C and Mande SS (2019) Web-gLV: A Web Based Platform for Lotka-Volterra Based Modeling and Simulation of Microbial Populations. Front. Microbiol. 10:288. doi: 10.3389/fmicb.2019.00288

INTRODUCTION

The ensemble of microbial groups residing in an ecosystem constitutes its microbiome. Mutual interactions between the resident microbes in a given microbiome depend not only on species diversity and abundances, but also on properties of their inhabited environment. On the other hand, the resident microbiota also has a profound influence on the properties of the habitat itself (Levy and Borenstein, 2013; Zelezniak et al., 2015). High throughput sequencing studies, especially for longitudinal microbiome projects, have greatly enhanced our understanding of the nature and dynamics of complex microbial interactions. Temporal analysis of microbial profiles has led to several intriguing findings (Gerber, 2014) and strengthened our understanding of the role of microbes in many diseases. Researchers have also reported new insights such as the existence of multiple steady states in human microbiome using time series microbiome experiments (Gajer et al., 2012; Faust et al., 2015).

Realizing the importance of the dynamic microbiome has encouraged development of methods and tools for its analysis and modeling (Fisher and Mehta, 2014; Bucci et al., 2016; Shaw et al., 2016; Baksi et al., 2018). Some of these tools provide specialized methods to visualize, cluster and compare temporally similar microbial groups, find causal relationships, analyze stationarity, identify

community-states, etc. Modeling microbial populations has recently attracted generous attention owing to its capability and potential to forecast future behaviors of the system as well as allow improved estimation of microbial interactions (Berry and Widder, 2014; Fisher and Mehta, 2014). The classical Lotka-Volterra equations can be used to model simple systems such as two species predator prey where the interactions are strictly assumed to be competitive. The "generalized" Lotka-Volterra (gLV) equations on the other hand are an extension of the logistic growth model and are more general than the classical predator-prey (Lotka-Volterra) equations where the interacting species might have a wide range of relationships including competition, cooperation, or neutralism. Such gLVs assume that the interaction (or the effect) of one species with another is encoded in the corresponding coefficient in the equation, providing a powerful framework to model and simulate microbial populations. It must be noted that gLV based models capture the interactions using a single averaged effect in a mean-field type model for which modest computational resource is sufficient. Consequently, it does not account for stochastic fluctuations (random processes), intrinsic dynamic correlations, and cannot address any emerging spatial structures which requires extensive computation. All the caveats applicable to extrapolation of the dynamics of a non-linear system apply to the predictions of the model. However, gLV formulations can still provide a reasonable starting point for more advanced community models and capture the effect of inter microbial associations in a more meaningful way as compared to conventional correlation based methods. Although correlations between groups of microbes can help in revealing underlying ecological processes, they are in most cases insufficient to serve as proxy for microbial interactions (Berry and Widder, 2014; Fisher and Mehta, 2014). Parameter estimations using Lotka-Volterra based models have been demonstrated to be better than correlation based measures (Fisher and Mehta, 2014). Additionally, the gLV models can provide an estimate of the native growth rates of uncultured microbes. While a positive value of the "interaction coefficient" is assumed to be a beneficial effect, a negative value indicates an inhibitory effect. If the coefficient has a zero value, no interaction is assumed to be present between the two taxa. The gLV equations were first used to model the interaction between bacteria and yeast in a cheese microbiome (Mounier et al., 2008) and thereafter in a few more microbiome studies (Marino et al., 2014; Dam et al., 2016; Vos et al., 2017; Venturelli et al., 2018). Simulation studies using generalized Lotka-Volterra (gLV) models can be used to understand microbiome dynamics and can assist biologists to design better experimental strategies. For a given microbial community (with known abundance and diversity), gLV can also be used to predict the future state of the microbiome. Similarly, it can be utilized to understand the temporal behavior of the microbiome if the initial conditions are perturbed.

Tools like LIMITS (Fisher and Mehta, 2014), MetaMis (Shaw et al., 2016), and MDSINE (Bucci et al., 2016) are available for applying gLV modeling on microbial time series data. LIMITS and MetaMis focus mainly on reconstruction of microbial interactions and are available as Mathematica code and an offline Matlab based GUI, respectively. MDSINE, although providing the most comprehensive suite of functionalities for analysis, requires knowledge of Matlab programming. In this communication, we present a web based tool called "Web-gLV" (freely available at http://web.rniapps.net/webglv) which can be used for modeling, visualization, and analysis of microbial populations without any programming expertise and has no installation requirements (**Supplementary Table 1**). Users can either upload a microbial time series abundance data matrix to formulate the mathematical models automatically or can provide pre-calculated model parameters, namely the growth rate, and inter-microbial interaction matrix. The outcomes of the simulations can be used to obtain various biological insights and enable optimization of experimental designs. "Web-gLV" is expected to be a valuable addition to the suite of tools in the field of ecological modeling and metagenomic systems biology.

RESULTS

"Web-gLV" provides an easy platform for biologists to exploit the benefits of gLV modeling by simply uploading the experimentally obtained time series microbial abundance data. The application is flexible to allow users input microbial growth rates and interaction values if known from other sources. We demonstrate the utility of "Web-gLV" using few publicly available datasets.

Case-Study 1: Predicting the Future State of Gut Microbiome

In this simulation, we used an available longitudinal metagenomic time series data of gut microbiome samples from a healthy human subject (Caporaso et al., 2011). The aim of this case study was to model the temporal behavior of top five dominant microbial taxa present in healthy human gut microbiome and use the model for predicting temporal dynamics of a future state which is unknown to the model. In order to achieve this, we used the above dataset to create a gLV model using the first 100 time points and considered the 101th time point as a start point to predict the abundance profiles of the subsequent 30 time points. The predicted 30 time points were then compared with the experimentally reported abundance profiles (**Supplementary Figure 1A**). In order to evaluate how close "Web-gLV" predicted trends are with respect to the experimentally observed trends, a Dynamic Time Warping (DTW) based algorithm (Berndt and Clifford, 1994) was used. DTW can evaluate the similarity between two time series of equal or unequal lengths using a dynamic programming based approach and can be used to successfully capture equivalence in the overall pattern. The low DTW distances between the observed and predicted trajectories (**Supplementary Figure 1B**) indicated that the gLV model was able to capture the observed temporal patterns in the selected taxonomic groups with good accuracy. The predicted dynamics could capture *Lachnospira*'s positive influence on *Faecalibacterium* as well as its negative influence on *Akkermansia*, *Bacteroides,* and *Phascolarctobacterium* (**Supplementary Figure 1C**). Cyclic trends in the *Bacteroides* abundance (as prevalent in the observed trends) were also seen

to be well captured in the predicted trajectories. In order to evaluate the robustness of the predicted trends, we changed the initial abundance values (to half and one fourth) of the two most abundant taxa, namely *Bacteroides,* and *Akkermansia*. With these changes, the predicted abundances of the two taxa did not show much deviation in their temporal trends (**Supplementary Figure 2**). Therefore, as expected, these two taxa, being the most abundant, were seen to be robust to different initial values. To check the similarities in the trends of the selected taxa over time, DTW distance metric was used to generate the dendograms (**Supplementary Figure 1D**). The obtained results indicated a good agreement between the observed and "Web-gLV" predicted trees, thereby validating the simulation capability of gLV models. The details of the individual steps followed in the case study are explained under the "Methods" section.

Case-Study 2: Understanding Changes in Microbial Interaction Patterns Upon Perturbation

The ability of gLV modeling to decipher interaction patterns in a microbial community was exploited in this case study to find differences between a healthy and perturbed gut microbiome. In order to understand the effect of perturbation on the dominant microbial genera, we used the publicly available time series microbiome data corresponding to *Clostridium difficile* infection (Bucci et al., 2016). The dataset consisted of regularly sampled time-series microbiome abundances (for 28 days) in five gnotobiotic mice pre-colonized with human commensal bacterial strains which were later infected with *C. difficile* spores. The data also included measured microbial abundances for additional 28 days post infection in these mice. For constructing the gLV model and predicting microbial interactions in the unperturbed state of the microbiome, we used the top five abundant taxa from data corresponding to the pre *C. difficile* infection time points of all five mice samples. Similarly, in order to construct a representative model and predict microbial interactions in the perturbed state, we considered the post *C. difficile* infection time points. Thus, two models, namely "normal state model" and "perturbed state model", were generated for each mouse sample using gLV modeling implemented in "Web-gLV." A biological realistic constraint enforcing positive intrinsic growth and negative or zero self interaction (Bucci et al., 2016) was considered during model generation (see Methods for details).

The predicted interaction profiles revealed a clear difference in the nature of microbial interactions between "normal" and "perturbed" states for all the five samples (**Supplementary Figure 3**). Upon inspecting the changes in the nature of interactions of individual taxa (from their normal to perturbed state) across all the samples, it was observed that genera exhibiting the maximum change differed in each of the samples (**Supplementary Figure 3**). For example, while *Akkermansia muciniphila* showed the least change across a majority of the mice samples ("Mouse 1," "Mouse 3" and "Mouse 4"), no clear cut pattern was observed for other taxa. Overall, the total negative interactions decreased in most of the samples ("Mouse 3," "Mouse 4" and "Mouse 5") but every mouse displayed a unique combination of interaction profiles. This may be explained as an effort by the dominant players (taxa) in the microbiome, each trying in its own way to influence the individual sample level variations.

In order to evaluate whether the model generated using the perturbed state of one mouse is able to predict the perturbation dynamics of the other mice, we predicted the post perturbation trajectories corresponding to each mouse, considering perturbation model of every other mice. Results of this simulation indicated an overall good prediction of the temporal dynamics (**Supplementary Figure 4**).The results corresponding to the predicted perturbed state trajectories of a mouse based on the model of its own normal state is shown (**Supplementary Figure 4**). To achieve this, gLV models were also generated for the normal states corresponding to all the mice samples (see Methods for details). In the next step, we checked whether these predictions could be improved by incorporating the normal state of the same mouse in combination with the perturbed state of another mouse. The comparison of the predicted and observed perturbation dynamics for each of the subjects was performed by evaluating their sum DTW distances. Comparison of the results indicated that the perturbed state of a mouse could be predicted better using the perturbed state of another mouse instead of using the model corresponding to the normal state of the same mouse. Interestingly, utilizing the normal state model of the same mouse in combination with the perturbed state model of another mouse did not show a consistent improvement (**Supplementary Figure 4**). Thus, the results indicate that the growth rate and interaction parameters of perturbation dynamics are better encoded in a comparable perturbation model rather than the normal state model of the same subject. However, additional advanced modeling steps like inclusion of antibiotic susceptibilities are expected to further enhance prediction accuracies (Bucci et al., 2016). The main objective of this case study was to demonstrate the capability of "Web-GLV" to use growth rate and interaction parameters derived from other experiments to perform simulations on new data in a user friendly way.

CONCLUSIONS

The increased affordability of DNA sequencing has enabled researchers to move beyond the hypothesis generated using static snapshots of microbiome. Lotka-Volterra based modeling provides an efficient means to leverage the current volume of generated longitudinal microbiome data. The generalized Lotka-Volterra (gLV) modeling extends the classical two species predator prey models which are widely used in ecology. An important advantage of gLV models is its ability to estimate the native growth and interaction parameters of uncultured microbes in a given environment from temporal data which would otherwise be difficult using traditional culture based methods (Bucci and Xavier, 2014). Consequently, using these parameters, one can study the changes in microbial communities over time starting with unknown initial conditions. This allows testing

of new hypothesis and helps to gather improved mechanistic insights. However, the intricacies of the advanced mathematical concepts involved as well as their implementation might prove to be a hindrance for many biologists. The limited availability of "ready-to-use" tools for such analysis also serves as a bottleneck to quickly test a hypothesis and obtain meaningful insights. We have developed "Web-gLV" to bridge this gap and to enable biologists take advantage of the multispecies modeling and simulation without any programming expertise. "Web-gLV" also bypasses any installation needs and requires only the time series microbial abundance data as input. A set of interactive operations allows easy initialization and simulation of microbial population as well as analysis of the output trajectories. Results of repeated simulations can be easily evaluated by altering the initial values and the parameters using GUI based inputs. The interactive graphical plots generated by the tool aids in easy analysis and comparison of the results. We demonstrate the ease with which Web-gLV can be used to automatically model and simulate microbial communities and generate outputs. Furthermore, we demonstrate the accuracy of the predictions and possible biological interpretations of the results.

Although gLV based models provide a good starting point for modeling microbial community dynamics, it does not account for random processes which forms essential part of any biological system. Additionally, with the increase in number of species and time span of prediction, the simulation output is also prone to numerical errors. Consequently, Web-gLV limits simulating a maximum of 10 species at a time for at the most 100 time points. The compositionality bias in microbiome data arising due to sampling and sequencing limitations may also cause inaccurate estimation of simulation parameters. Moreover, too much irregularity in the sampled time points may also result in inaccurate parameter estimations. Hence, it is advised to cautiously interpret the findings obtained using Web-gLV and more importantly augment it with the underlying biology of the systems (Faust and Raes, 2012; Gerber, 2014).

MATERIALS, METHODS, AND IMPLEMENTATION

Modeling and Parameter Estimation of a Generalized Lotka-Volterra Equation

A multi species gLV model for rate of change of a counts "x_i" of a species "i" can be written as an ordinary differential equation as shown below:

$$\frac{dx_i}{dt} = x_i \left(r_i + \sum_{j=1}^{n} \propto_{ij} x_j \right) \ldots \quad (1)$$

where, r_i corresponds to the intrinsic specific growth rate of species "i" and $\propto_{ij}$ is the influence on the growth rate of species "i" exerted by another species "j" of the community consisting of "n" species. Thus, for a given set of "n" species, "n" differential equations can be formulated which can then be used for simulating the behavior of those species starting with a set of initial values. However, in order to perform such simulation, one also needs to find the values of other types of parameters for each of the equations namely the growth rate r_i and the set of inter-species interaction parameter $\propto_{ij}$.

The Equation (1) can be rewritten as below:

$$\frac{1}{x_i}\frac{dx_i}{dt} = \left(\mathbf{r_i} + \sum_{\mathbf{j=1}}^{\mathbf{n}} \propto_{\mathbf{ij}} \mathbf{x_j} \right) \ldots \quad (2)$$

Further, Equation (2) can be expressed as:

$$\frac{d\ln(x_i\,(t))}{dt} = \left(r_i + \sum_{j=1}^{n} \propto_{ij} x_j \right) \ldots \quad (3)$$

For numerical integration following the implicit trapezoid method, upon discretizing Equation (3) for each sub interval (let $[k,k+1]$), and taking the average value of x_j we get:

$$\ln x_i\,(t_{k+1}) - \ln x_i\,(t_k) \approx \left(r_i + \sum_{j=1}^{n} \propto_{ij} \left\{ \frac{\left(x_{j(k+1)} + x_{jk} \right)}{2} \right\} \right) \Delta t \quad \ldots \quad (4)$$

Now, given a time series data for abundances of the set of "n" species, these two parameters namely r_i and $\propto_{ij}$ can be estimated by comparing equation (4) to a linear regression model for log lagged differences in abundances estimated for each i^{th} taxa (x_i) available in the microbiome time series data wherein the intercept corresponds to the r_i values and the coefficients to the $\propto_{ij}$ values (**Figure 1**). Earlier studies have suggested using a constrained regression (with enforced positive intrinsic growth and negative or zero self interaction constraints) for microbial populations as it is biologically more realistic (Bucci et al., 2016). Web-gLV implements two methods for parameter estimation namely PLSR (Partial least squares regression) for unconstrained estimation (Haenlein and Kaplan, 2004) and LSEI algorithm (Haskell and Hanson, 1978) for constrained estimations (**Supplementary Figure 5**). The constrained estimation solves a least square problem under conditions where r_i is forced to take a positive value and $\propto_{ii}$ values are constrained to less than or equal to zero.

Evaluation of Predicted Trajectory

The observed and predicted trajectories are compared using a Dynamic Time Warping algorithm (DTW). DTW measures the similarity between two time series (with or without a lag) using a dynamic programming approach (Berndt and Clifford, 1994) and can be used to compare time series of unequal lengths. As, in most cases, the compared trajectories in "Web-gLV" are expected to be unequal, DTW fits as the best scoring metric. If "T1" and "T2" are two time series vectors of length "m" and "n," respectively, DTW finds a mapping path $\{(p_1,q_1),(p_2,q_2),\ldots,(p_k,q_k)\}$ with boundary conditions $(p1,q1)=(1,1)$ and $(p_k,q_k)=(m,n)$. The DTW distance

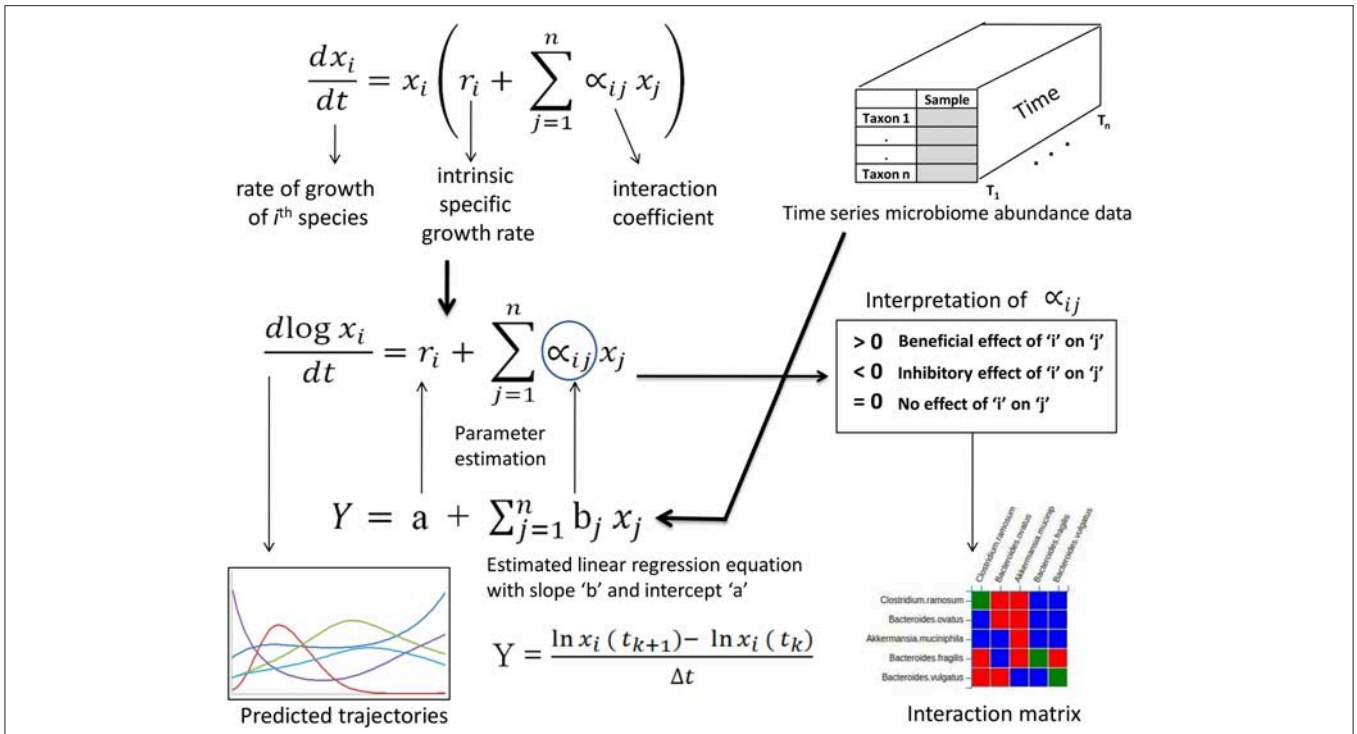


FIGURE 1 | Overview of formulation and use of multi species generalized Lotka-Volterra (gLV) models for obtaining microbial interactions and predict future trajectories.

between T1 and T2 for a point (*i, j*) is calculated by solving a dynamic programming using the distance formula shown below:

$$DTW\left(i,j\right) = \left|T1\left(i\right) - \ T2(j)\right| + min\left\{\begin{matrix} DTW\left(i-1\right),j \\ DTW(i-1,j-1) \\ DTW(i,j-1) \end{matrix}\right\} \ldots(5)$$

To calculate the final distance, a matrix MDTW of dimensions m×n is constructed after filling MDTW(1,1) with the initial condition value of MDTW(1,1) = $\left|T1\left(1\right) - \ T2(1)\right|$. The whole matrix is then filled one element at a time using the formula shown in Equation 5. The final distance value is available at the cell MDTW(*m,n*). The distance is calculated between the scaled (between 0 and 1) time series belonging to the "Observed" and "Predicted" data which is presented as a table along with the trend plots in the "Web-gLV" tool. The sum total (or cumulative) DTW distance for a set of predicted taxa can be used as a measure to score the similarity between two or more simulations. Additionally, the "all vs. all" DTW distance is calculated for the "Observed" and "Predicted" data to generate the hierarchically clustered dendograms. These dendograms can be useful to understand the microbial community structure.

Implementation of the Web-gLV Tool

Web-gLV has been developed using JavaScript (and PHP) for the frontend with R (deSolve package) and Perl scripts in the backend (Soetaert et al., 2010). The tool can perform simulations starting two types of input sets. A user can either upload only a taxonomic abundance file which will be used to estimate parameters and generate reference plots for the observed trends. Alternatively, in addition to a taxonomic abundance file, a growth rate file and inter-taxa interaction file can also be uploaded separately to bypass the automatic parameter estimation step and use the supplied values for numerical simulation. A metadata file corresponding to the timepoints specified in the main taxonomic abundance file can also be uploaded as an optional input. This metadata information will be used by the tool to augment the time series plots based on the available information. The reference values of initial starting point of simulation for the selected taxa set can be selected from one of the time point row of this abundance table. Once the input files are uploaded, the various steps involved in running a simulation are described below:

Step 1: Selecting the taxa required for simulation from the input dataset:

Given a time series microbiome data as input, the tool presents a tabulated graphical summary in the form of box plots, trend charts and other accompanying statistics of the input microbial abundance profiles (**Figure 2A**). Additionally, a Pearson correlation ($r \geq 0.5$ and $r \leq -0.5$) based network is created using the core taxa (having <30% zeroes in the sampled longitudinal timescale) (**Figure 2B**). This network can be viewed by clicking on the link "Click here to show/hide correlation network." The taxonomic groups desired to be added for modeling can be selected using the graphical summary table, the dropdown search box or the correlation network. Clicking on a taxa label in the summary table adds that taxa to the simulation. Similarly, clicking on a node in the network adds

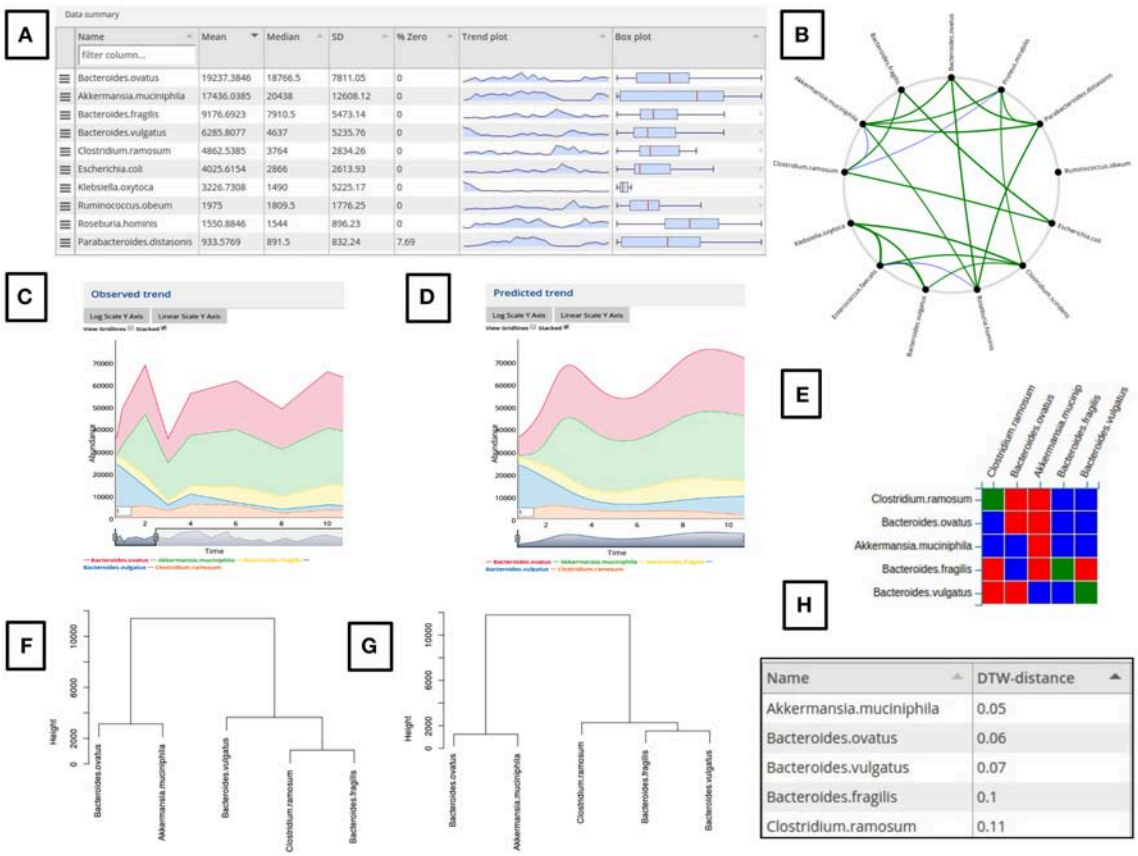


FIGURE 2 | Demonstration of the various features of the web-gLV tool. **(A)** Tabulated summary of the input microbiome abundance data. **(B)** Microbial association network generated from the input data which can be used to select the required taxa to be used for modeling. **(C)** A stacked line plot based comparison of the observed **(C)** and predicted **(D)** trajectories. **(E)** A matrix representation of the predicted interaction coefficients for the modeled taxa (Red, Negative, Blue, Positive and Green, No interaction). Dendogram based comparison of the change in the microbial community structure between the observed **(F)** and the predicted **(G)** trends. **(H)** Evaluation of the similarity between the observed and predicted time series curves scored using a DTW metric.

it and the connected nodes. This feature can be used to select a set of closely related microbial groups showing a correlated temporal behavior. Adding a taxa for simulation using the above two methods also makes it visible in the searchable dropdown along with a graphical display of its temporal behavior in the "Observed trend" window. This dropdown can also be used to remove added taxa or add more taxa by selecting from the dropdown. Adding or deleting a taxon automatically updates the "Observed trend" plot. Several user interactive operations like log transformation, stacking/un-stacking, viewing gridlines and selecting a desired window of the trend plot is possible. A moving average based smoothing can also be applied to the time series plot by modifying the value in the left bottom corner box (**Figures 2C,D**).

Step 2: Selecting simulation parameters:

After selecting the taxonomic groups, a user has to specify the modeling parameters like start and end point of data time-points for estimating the interaction coefficients, numerical simulation interval duration and the solver used for numerical integration of the ordinary differential equations (ODE) method. The interaction coefficients for the equation are then inferred using a partial least square regression (if selected for unconstrained growth rate selection) or a constrained regression (if selected for an enforced positive intrinsic growth and negative/zero self interaction constraints). Other parameter estimation methods that require numerical integration at each step of the optimization process are potentially better in terms of accuracy but require substantially more resources and time than the implemented methods. Earlier studies have suggested using the constrained method for modeling microbial populations as it is biologically more realistic (Bucci et al., 2016). The start time (or initial value) for the simulation can be interactively selected as any one of the time-point from the input dataset with provision to edit the values. This option can be used to test perturbations in the initial microbial abundance values and observe the simulation outcomes. It needs to be noted that the "Parameter estimation" settings are not available when a simulation is started with a user supplied growth rate and interaction file.

Step 3a: Running the simulation:

After setting the parameters, The "Run simulation" button can be clicked to perform a simulation. If the simulation is successful, the predicted trajectories for the selected taxa are displayed under the "Predicted trend" window (**Figure 2D**). The observed vs. predicted trend plot for a taxon is also generated as a mixed plot with the observed trends shown in points connected by dotted lines and the predicted as firm lines of same color. In case of an unsuccessful simulation due to incorrect parameter or solver limitation, an error message is displayed and no trajectories are generated. The timeseries plot in the "Observed trend" window is automatically set to display the selected time range if the simulation range matches. This feature is helpful to compare the predicted trajectories from a modified starting point and compare it with the unmodified observed trends. The predicted growth rate and interaction coefficient matrix (**Figure 2E**) which was used for simulation is displayed graphically for convenience. A simulation can be re-run by altering some parameter/simulation settings as well as with a modified set of initial values.

Step 3b: Performing cross predictions:

"Web-gLV" can also be used to perform cross predictions by estimating growth and interaction parameters in one simulation and use the same to predict dynamics in a different simulation. The predicted parameters can be saved as text files using the "Download table" option available under "Predicted Intrinsic Growth Rates" and "Predicted Interaction Matrix" headers in the "Web-gLV" tool. While performing a new simulation with a similar type of taxonomic groups whose time series abundances are available, the downloaded parameters can be uploaded to perform the simulation. This feature available in the "Web-gLV" tool can be used to test the prediction performance of models on unknown initial conditions as demonstrated in case study 2.

Step 4: Evaluating the simulation output:

The predicted trajectories are scored for their similarities (**Figure 2H**) with the observed time series using a Dynamic Time Warping (DTW) distance metric (Berndt and Clifford, 1994). The all vs. all DTW metric is used to construct a hierarchal clustered dendogram for the observed and predicted trends (**Figures 2F,G**). These dendograms represents the temporal similarities between the selected microbial groups and hence a reflection of their community structure. A comparison of the dendograms generated for the "Observed" and "Predicted" data can hence be used as a measure of the simulation prediction accuracy.

Numerical Validation of Web-gLV Predictions

Web-gLV implements two methods for parameter estimation namely PLSR (Partial least squares regression) for unconstrained estimation (Haenlein and Kaplan, 2004) and LSEI algorithm (Haskell and Hanson, 1978) for constrained estimations. We used standard R modules namely pls and limSolve, respectively, for the backend implementation. The tool is designed to capture trends, which provides an idea of the growth rate and nature of interactions. However, for an improved understanding, it is imperative to look into the functional potential of the participating taxonomic groups (Nagpal et al., 2016; Bhatt et al., 2018). Web-gLV can provide a good starting point for more advanced community models by augmenting information from other sources. We compared both the constrained as well as unconstrained parameters estimated by web-gLV with previously reported methods as demonstrated in section introduction of **Data Sheet 1**. It should be noted that the calculated coefficients for the constraint optimization solves the same problem in different ways providing non-unique solutions. Consequently, the parameters are free to take any values depending on the solution which may result in differences between the estimated parameter values. However, as expected, the predicted trajectories (when evaluated for the case studies) show a good agreement between the various tools (Section results of **Data Sheet 1**).

Using "Web-gLV" to Perform the Case Studies

The modeling and simulations involved in the case studies demonstrated in the "Results" section were performed completely using the "Web-gLV" tool. The datasets are available in the home page of the tool which can be auto-loaded by selecting the "View" button corresponding to each case study. The first 100 time point for case study 1 were selected using Timepoint 1 (sampling interval: 0) as start and Timepoint 100 (sampling interval: 143) as end under the "parameter estimation settings." The future 30 time points were predicted by selecting Timepoint 101 (sampling interval: 144) as start and setting the "Time duration" option to 30 under "simulation settings." For case study 2, the start and end time points for creating the "normal" state models corresponded to Timepoint 1 (sampling interval: 0.75) and Timepoint 13 (sampling interval: 28), respectively. Similarly, for the perturbed models, Timepoint 1 (sampling interval: 28.75) and Timepoint 26 (sampling interval: 56) corresponded to the start and end time points, respectively. The solver for numerical simulation was selected as ODE45 for both the case studies with time interval as 0.1. A biological realistic constraint enforcing positive intrinsic growth and negative or zero self interaction was applied for generating all the modes by selecting the option under "parameter estimation settings." However, the constrained parameter optimization failed to find an exact solution for the "normal" state model of "Mouse 5" for which we unselected the option and generated the model without the constraints.

DATA AVAILABILITY

The link https://web.rniapps.net/webglv contains the data-sets used in the case study along with the user manual for running the Web-gLV tool.

AUTHOR CONTRIBUTIONS

BK conceived the idea, implemented the algorithms and developed the interface. BK, CG, and SM designed the case studies, evaluated the results, and drafted the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

BK is an industry sponsored AcSIR Ph.D. student at Chemical Engineering & Process Development Division, CSIR-National Chemical Laboratory (NCL), Pune 411008 (India), and would like to acknowledge NCL and the Academy of Scientific and Innovative Research (AcSIR) for its support. CG acknowledges funding from DST, India through SERB grant EMR/2017/003271.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00288/full#supplementary-material

REFERENCES

Baksi, K. D., Kuntal, B. K., and Mande, S. S. (2018). "TIME": a web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Front. Microbiol.* 9:36. doi: 10.3389/fmicb.2018.00036

Berndt, D. J., and Clifford, J. (1994). "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* AAAIWS'94 (Seattle, WA: AAAI Press), 359–370. Available online at: http://dl.acm.org/citation.cfm?id=3000850.3000887 (Accessed May 14, 2018).

Berry, D., and Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* 5:219. doi: 10.3389/fmicb.2014.00219

Bhatt, V., Mohapatra, A., Anand, S., Kuntal, B. K., and Mande, S. S. (2018). FLIM-MAP: gene context based identification of functional modules in bacterial metabolic pathways. *Front. Microbiol.* 9:2183. doi: 10.3389/fmicb.2018.02183

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

Bucci, V., and Xavier, J. B. (2014). Towards predictive models of the human gut microbiome. *J. Mol. Biol.* 426, 3907–3916. doi: 10.1016/j.jmb.2014.03.017

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011). Moving pictures of the human microbiome. *Genome Biol.* 12:R50. doi: 10.1186/gb-2011-12-5-r50

Dam, P., Fonseca, L. L., Konstantinidis, K. T., and Voit, E. O. (2016). Dynamic models of the complex microbial metapopulation of lake mendota. *npj Syst. Biol. Appl.* 2:16007. doi: 10.1038/npjsba.2016.7

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66. doi: 10.1016/j.mib.2015.04.004

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550. doi: 10.1038/nrmicro2832

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:e102451. doi: 10.1371/journal.pone.0102451

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., et al. (2012). Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* 4:132ra52. doi: 10.1126/scitranslmed.3003605

Gerber, G. K. (2014). The dynamic microbiome. *FEBS Lett.* 588, 4131–4139. doi: 10.1016/j.febslet.2014.02.037

Haenlein, M., and Kaplan, A. M. (2004). A Beginner's guide to partial least squares analysis. *Understand. Stat.* 3, 283–297. doi: 10.1207/s15328031us0304_4

Haskell, K. H., and Hanson, R. J. (1978). *Algorithm for Linear Least Squares Problems With Equality and Nonnegativity Constraints.* Albuquerque, NM: Sandia Labs. doi: 10.2172/6461744

Levy, R., and Borenstein, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl. Acad. Sci. U.S.A.* 110, 12804–12809. doi: 10.1073/pnas.1300926110

Marino, S., Baxter, N. T., Huffnagle, G. B., Petrosino, J. F., and Schloss, P. D. (2014). Mathematical modeling of primary succession of murine intestinal microbiota. *PNAS* 111, 439–444. doi: 10.1073/pnas.1311322111

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

Nagpal, S., Haque, M. M., and Mande, S. S. (2016). Vikodak - a modular framework for inferring functional potential of microbial communities from 16S metagenomic datasets. *PLoS ONE* 11:e0148347. doi: 10.1371/journal.pone.0148347

Shaw, G. T., Pao, Y.-Y., and Wang, D. (2016). MetaMIS: a metagenomic microbial interaction simulator based on microbial community profiles. *BMC Bioinformatics* 17:488. doi: 10.1186/s12859-016-1359-0

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010). Solving differential equations in R: package deSolve. *J. Stat. Softw.* 33, 1–25. doi: 10.18637/jss.v033.i09

Venturelli, O. S., Carr, A. C., 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

Vos, M. G. J., de, Zagorski, M., McNally, A., and Bollenbach, T. (2017). Interaction networks, ecological stability, and collective antibiotic tolerance in polymicrobial infections. *PNAS* 114, 10666–10671. doi: 10.1073/pnas.1713372114

Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D. R., Bork, P., and Patil, K. R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl. Acad. Sci. U.S.A.* 112, 6449–6454. doi: 10.1073/pnas.1421834112

Conflict of Interest Statement: BK and SM are employed by the company Tata Consultancy Services Limited.

The remaining 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.

Copyright © 2019 Kuntal, Gadgil and Mande. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 22 February 2019
doi: 10.3389/fmicb.2019.00118

Global Microbiome Diversity Scaling in Hot Springs With DAR (Diversity-Area Relationship) Profiles

Lianwei Li [1,2] *and Zhanshan Ma* [1,2,3]*

[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] Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, China, [3] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

The spatial distribution of biodiversity (i.e., the biogeography) of the hot-spring microbiome is critical for understanding the microbial ecosystems in hot springs. We investigated the microbiome diversity scaling (changes) over space by analyzing the diversity-area relationship (DAR), which is an extension to classic SAR (species-area relationship) law in biogeography. We built DAR models for archaea and bacteria with 16S-rRNA sequencing datasets from 165 hot springs globally. From the DAR models, we sketch out the biogeographic maps of hot-spring microbiomes by constructing: (i) DAR profile—measuring the archaea or bacteria diversity scaling over space (areas); (ii) PDO (pair-wise diversity overlap or similarity) profile—estimating the PDO between two hot springs; (iii) MAD (maximal accrual diversity) profile—predicting the global MAD; (iv) LRD/LGD (ratio of local diversity to regional or global diversity) profile. We further investigated the differences between archaea and bacteria in their biogeographic maps. For example, the comparison of DAR-profile maps revealed that the archaea diversity is more heterogeneous (i.e., more diverse) or scaling faster than the bacterial diversity does in terms of species numbers (species richness), but is less heterogeneous (i.e., less diverse) or scaling slower than bacteria when the diversity (Hill numbers) were weighted in favor of more abundant dominant species. When the diversity is weighted equally in terms of species abundances, archaea, and bacteria are equally heterogeneous over space or scaling at the same rate. Finally, unified DAR models (maps) were built with the combined datasets of archaea and bacteria.

Keywords: biogeography of hot-spring microbiome, DAR (diversity-area relationship), MAD (maximal accrual diversity), local to regional (global) diversity (LED/LGD), biogeographic differences between archaea and bacteria

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Long Gao,
University of Pennsylvania, United States
Sushant Patil,
University of Chicago, 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: *17 October 2018*
Accepted: *18 January 2019*
Published: *22 February 2019*

Citation:
Li L and Ma Z (2019) Global Microbiome Diversity Scaling in Hot Springs With DAR (Diversity-Area Relationship) Profiles. Front. Microbiol. 10:118. doi: 10.3389/fmicb.2019.00118

INTRODUCTION

Hot springs are one of the extreme environments on the earth planet. Hot spring microbiomes play a critical role in shaping the geothermal ecosystems. The structures and functions of microbial communities inhabiting hot springs have their somewhat unique characteristics compared with non-geothermal environment microbiomes (Inskeep et al., 2013a,b; Song et al., 2013; Masaki et al., 2016; Poddar and Das, 2018), such as soil microbiome (Fierer et al., 2012; Hartmann et al., 2014), marine microbiome (Gajigan et al., 2018), and human microbiome (Huttenhower et al., 2012). Hot springs are often abundant in thermophilic, hyperthermophilic, and thermoresistant bacterial

and archaeal taxa (Urbieta et al., 2014a,b). For example, the hot springs located in the Chilas and Hunza areas of Pakistan with 90–95°C temperature harbor abundant phylum *Thermotogae* (Amin et al., 2017). Many reports have suggested that hot spring microbial communities are extremely heterogeneous and are often dominated by the thermophilic bacterium (e.g., Cole et al., 2013; Sharp et al., 2014; Masaki et al., 2016; Poddar and Das, 2018). Even though some recent studies showed that microbial eukaryotes, especially microbes from the phyla of *Ascomycota* and *Basidiomycota*, may also be important components of hot springs microbiomes (de Oliveira et al., 2015; Salano et al., 2017; Liu et al., 2018; Oliverio et al., 2018), the microbial communities routinely consist of Bacteria and Archaea (Meyer-Dombard and Amend, 2014; Hedlund et al., 2015; Merkel et al., 2017).

The stability of hot spring environments is routinely determined by the steady state of their microbial diversity in a specific environment, where water temperature, pH, and chemical composition are often the most important factors to influence the diversity (Mohanrao et al., 2016; Amin et al., 2017; Chan et al., 2017; Ghilamicael et al., 2017; Poddar and Das, 2018). In general, there is an inversely proportional relationship between microbial diversity and temperature of hot spring (Cole et al., 2013; De León et al., 2013; Amin et al., 2017; Chan et al., 2017). In addition to temperature, water pH is another primary environmental factor directly influencing microbial diversity in hot springs (Inskeep et al., 2013a,b; Xie et al., 2015). The water pH is determined by the chemical composition in hot springs. The role of chemical composition in designing the structure of geothermal microbial communities should not be underestimated, which sometimes play together with the pH (Jiang et al., 2015; Geesey et al., 2016). In spite of the extensive studies on the microbial diversities in hot springs, which suggest that the biodiversity of microbial communities vary with physicochemical conditions and biogeographical location of inhabiting hot springs, to the best of our knowledge, the hot-spring microbiome diversity scaling on regional or global scales from a biogeography perspective has not been investigated yet.

To investigate the microbiome diversity scaling on regional/global scale, one of the most powerful theoretical tools is the classic species-area relationship (SAR) power law, which achieved a rare "law" status in ecology and biogeography. The SAR is often described with a power function $S = cA^z$, where S is the number of species accumulated from a region of size A, and z is termed species (number) scaling parameter. The study of SAR can be traced back to the nineteenth century (Watson, 1835; Arrhenius, 1921; Preston, 1960, 1962) and the relationship was said to inspire MacArthur and Wilson (1967) to establish their island biogeography theory, which helped to shift the focus of ecology from population level to community level.

Theoretically, a series of extensions of the classic SAR to general diversity-area relationship (DAR) were introduced by the author's group recently (Ma, 2018a,b). The extensions were justified to remedy a limitation with classic SAR, where the biodiversity is measured with the number of species or the so-termed *species richness*. Species richness can be a rather meaningful measure for biodiversity in the case of large plants and animals, but in many other cases (especially for microbes), it is a poor measure of biodiversity because it ignores the differences in species abundances. For example, 1,000 of panda and one billion of panda will weigh in the same with species richness; but if the latter number were the case, panda would not have been on the endangered species list. The DAR extension was facilitated by the adoption of the Hill numbers (Hill, 1973; Chao et al., 2012, 2014) as general diversity measures, which weight diversity differently depending on the so-termed diversity order. In terms of the Hill numbers, biodiversity can be measured by the so-termed diversity profile (Chao et al., 2012, 2014), which calculates a series of Hill numbers, weighted differently by the species abundance distribution (SAD). Therefore, Hill numbers are now well-recognized as the most appropriate measures for alpha-diversity and its multiplicative partition of beta-diversity is also considered with advantages over other beta-diversity measures. With the new DAR approach, four sets of new tools: DAR profile, PDO (pair-wise diversity overlap) profile, and MAD (maximal accrual diversity) profile, and LRR/LGR [local to regional (global) diversity ratio], can be established with the parameters from DAR modeling. These profiles, together with DAR models offer powerful tools, not only for quantifying the regional/global scaling of biodiversity, but also for sketching out the biogeography maps of biodiversity distribution (Ma, 2018a,b). In the present study, we apply the DAR approach to analyzing the global biodiversity scaling of hot-spring microbiome by reanalyzing the 16S-rRNA marker gene abundance datasets of 165 hot springs on a global scale, previously collected and published by Sharp et al. (2014). We further sketch out and compare the biogeography "maps" of archaea and bacteria, and highlight their differences in biodiversity distribution.

The four profiles, DAR (diversity-area relationship), PDO (pair-wise diversity overlap), MAD (maximal accrual diversity) profile, LRD (local to regional diversity ratio), we build for archaea, bacteria, and their combined assemblages offer tools to sketch out the biogeography maps with different themes. While, the map theme profiled by DAR is the diversity scaling (difference or heterogeneity) over space, the theme profiled by PDO is the similarity of diversity over space. While the map theme profiled by MAD is the theoretically maximal accrual diversity (essentially the maximal gamma diversity), the theme profiled by LRD/LGD is then the local *vs*. regional diversity comparison, which answers a simple question: how much, on average, a local sample can represent the regional or global diversity.

MATERIALS AND METHODS

The Hot Spring Microbiome Datasets

The datasets of 165 hot-spring microbiome were originally collected and reported by Sharp et al. (2014). Their 16S-rRNA OTU (operational taxonomic unit) tables were generated from 165 microbiome samples taken from sediment, soil, and mat in Western Canada and Taupo Volcanic Zone, New Zealand (Sharp et al., 2014). A total of 1,162,553 high quality sequences were obtained from the 165

samples with 634–15,283 sequences per sample. There were 61,910 OTUs, including 7,964 archaea and 53,946 bacteria, when those sequences were clustered at the 97% identity threshold. Further information on the datasets is referred to Sharp et al. (2014).

Definitions and Computational Procedures

Three steps are involved in building DAR models with microbiome datasets (see **Figure 1**): (i) bioinformatics analysis of 16S-rRNA data to get OTU tables (e.g., Schloss et al., 2009; Caporaso et al., 2010; Sinha et al., 2015). The microbiome quality control project: baseline study design and future directions. Genome Biology. Vol. 16: 276, https://doi.org/10.1186/s13059-015-0841-8; (ii) computing species or OTU diversities measured with the Hill numbers (Chao et al., 2012, 2014; Ma, 2017); (iii) constructing the DAR models (Ma, 2018a,b).

Diversity Measured in Hill Numbers

The Hill numbers (Hill, 1973; Jost, 2007; Chao et al., 2012) are considered as the most appropriate measure for alpha diversity,

$$ {}^{q}D = \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)} \tag{1} $$

where S is the number of species, p_i is the relative abundance of species i, q is the order number of diversity. The Hill number is not defined when $q = 1$, but its limit as q approaches to *1* exists in the following form:

$$ {}^{1}D = \lim_{q \to 1} {}^{q}D = \exp\left(-\sum_{i=1}^{s} p_i \log(p_1) \right) \tag{2} $$

The parameter q determines the sensitivity of the Hill number to the relative frequencies of species abundances. If $q = 0$, the species abundances do not weigh at all and ${}^{0}D = S$, which is simply the species richness. When $q = 1$, ${}^{1}D$ equal the *exponential* of Shannon entropy, and is interpreted as the number of typical or common species in the community because ${}^{1}D$ is weighted proportionally by species abundances. When $q = 2$, ${}^{2}D$ equal the reciprocal of Simpson index, i.e., the number of *dominant* or very abundant species in the community (Chao et al., 2012) because ${}^{2}D$ is weighted in favor of more abundant species. The general interpretation of ${}^{q}D$ (diversity of order q) is that the community has a diversity of order q, which is equivalent to the diversity of a community with ${}^{q}D = x$ equally abundant species. The so-termed diversity profile refers to the Hill numbers at different diversity order q (Jost, 2007; Chao et al., 2012, 2014).

The DAR Models and DAR Profile

Ma (2018a) extended SAR (species area relationship) to general DAR (diversity area relationship), in which diversity is measured with Hill numbers. The first DAR model, which borrowed the same power law (PL) function from the classic SAR, is:

$$ {}^{q}D = cA^{z} \tag{3} $$

where ${}^{q}D$ is diversity measured in the *q-th* order Hill numbers, A is *area*, and c and z are parameters.

A second DAR model is the power law with exponential cutoff (PLEC) model, which was originally introduced to SAR modeling by Plotkin et al. (2000), Ulrich and Buszko (2003), and Tjørve (2009) is:

$$ {}^{q}D = cA^{z} \exp(dA), \tag{4} $$

where d is a third parameter and is usually negative in the DAR models, and $\exp(dA)$ is the exponential decay term, which eventually overwhelms the power law behavior at very large value of A.

The following log-linear transformed equations can be used to estimate the parameters of the DAR models:

$$ \ln(D) = \ln(c) + z\ln(A) \tag{5} $$

$$ \ln(D) = \ln(c) + z\ln(A) + dA \tag{6} $$

The linear correlation coefficient (R) and p-value are used to judge the goodness of the model fitting.

Ma (2018a) defined the relationship between DAR-PL (power law) model parameter (z) and diversity order (q), or z-q trend, as the *DAR profile*, which comprehensively describes the change of diversity scaling parameter (z) with the diversity order (q).

Predicting MAD (Maximal Accrual Diversity) With PLEC-DAR Models

Ma (2018a) derived the maximal accrual diversity (MAD) in a cohort or population based on PLEC model [Equations (4) and (6)] as follows:

$$ Max({}^{q}D) = {}^{q}D_{\max} = c(-\frac{z}{d})^{z} \exp(-z) = cA_{\max}^{z} \exp(-z) \tag{7} $$

where $A_{\max}$ is the number of areas accrued to reach the maximum and is equal to:

$$ A_{\max} = -z/d \tag{8} $$

and all parameters are the same as in Equations (4) and (6).

Similar to the previous definition for DAR profile (z-q pattern), Ma (2018a) defined the MAD profile ($D_{\max}$-q pattern) as a series of $D_{\max}$ values corresponding to different diversity order (q).

Pair-Wise Diversity Overlap (PDO) Profile

The *pair-wise diversity overlap* (g) of two bordering areas of the same size (i.e., the proportion of the new diversity in the second area) is:

$$ g = 2 - 2^{z} \tag{9} $$

where z is the scaling parameter of DAR-PL model [Equations (3) and (5)]. If $z = 1$, then $g = 0$ and there is no overlap (similarity);

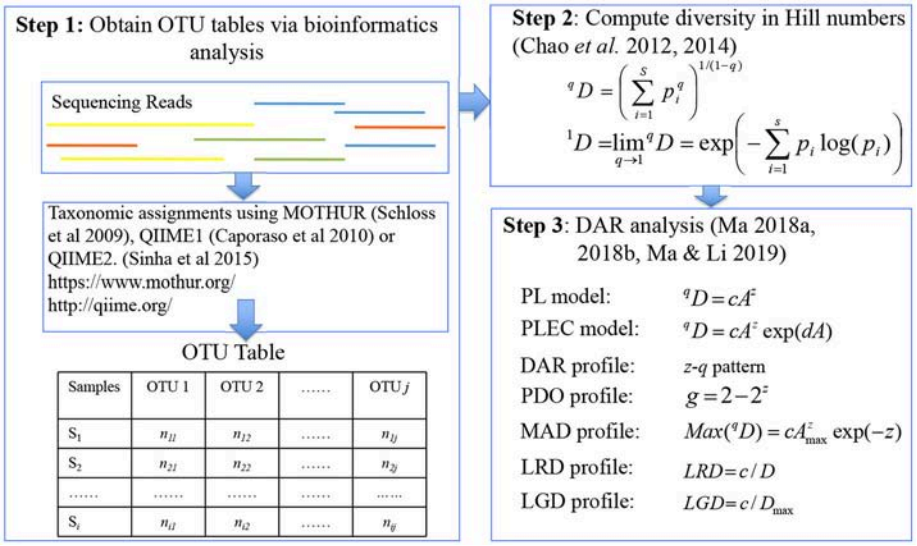


FIGURE 1 | A diagram showing the three steps involved in DAR (diversity-area relationship) analysis: major software packages (MOTHUR, QIIME, QIIME2), data format (OTU table), definitions (Hill numbers), models (PL and PLEC), and concepts (profiles) were illustrated.

and if $z = 0$, then $g = 1$ and there is a total overlap. In reality, g should be between 0 and 1. Since g is between 0 and 1, one may even use percentage notation to measure PDO.

Similar to previous definitions for DAR profile (z-q pattern) and MAD profile ($D_{\max}$-q pattern), Ma (2018a) defined the PDO profile (g-q pattern) as a series of PDO values corresponding to different diversity order (q).

The Ratio of Local Diversity to Regional (or Global) Accrual Diversity

Ma and Li (2019) defined the LRD (or LGD) as the ratio of local diversity of an *averaged area* to the *regional diversity* accrued or the global MAD (maximal accrual diversity). The dividend (local diversity) is ideally estimated with the parameter c of the DAR-PL model, but can be approximated with the parameter c of the DAR-PLEC model. The divisor can be either regional accrual diversity (which can be estimated with PLEC model directly) or global maximal accrual diversity (which is simply the MAD or D_{max}). Hence, in general, two similar metrics can be defined, depending on the regional or global scale is adopted: one is the ratio of local diversity to regional diversity (LRD), and another is the ratio of local diversity to global MAD (LGD). The LRD (LGD) can be computed with the following formulae, respectively:

$$LRD = c/D \tag{10a}$$

$$LGD = c/D_{\max} \tag{10b}$$

where D can D_{max} be computed with the PLEC model directly (Equations 4 and 7, respectively), c can be estimated or approximately with the PL or PLEC model. The LRD (LGD) at different diversity orders (q = 0, 1, 2, …) were defined as LRD (LGD) profile, or local to regional (local to global) diversity scaling profile (Ma and Li, 2019). It is essentially the ratio of alpha to gamma diversity.

Re-sampling Procedure to Enhance the Robustness of DAR Modeling

The accumulation order of areas in DAR modeling may influence the estimation of parameter c in fitting PL/PLEC models (Equations 3–6). When there is not a natural spatial sequence (or arrangement) among the communities sampled, or the arrangement information is not available, arbitrarily choosing an accumulation order (arrangement) can be problematic. To avoid the potential bias from an arbitrary order of the hot spring microbiome samples, we totally permutated the orders of all the community samples under investigation, and then randomly choose 100 orders of the communities generated from the permutation operation. In other words, rather than taking a single arbitrary order for accruing community samples in one-time fitting to the DAR model, we iteratively perform the DAR model-fitting 100 times with the 100 randomly chosen orders. Finally, the averages of the model parameters from the 100 times of DAR fittings are adopted as the model parameters of the DAR. In the case of this study, we do not have detailed information on the geographic locations of the hot-spring microbiome samples, the re-sampling scheme is adopted to remedy the deficiency.

RESULTS AND DISCUSSION

DAR Analysis of the Archaea

Table 1 displays the parameters from fitting the DAR (diversity-area relationship) PL (power law) and PLEC (power law with exponential cutoff) models to the datasets of archaea in hot springs. The p-values in **Table 1** show statistically significant fitting (in all models $p < 0.001$) of both PL and PLEC to

the datasets. The PLEC model has an advantage of being able to estimate the MAD (maximal accrual diversity) or D_{max}, which essentially measures the accrued diversity in a population or cohort, with the so-termed MAD-profile, as explained previously. The other two parameters from PL model, scaling parameter (z) and pair-wise diversity overlap (PDO) parameter (g) define the DAR-profile and PDO profile, respectively. From **Table 1**, we summarized the following findings:

(i) The archaea diversity scaling (changes across space) fitted to the PL model successfully, with $p < 0.001$, and succeeded in all 100 times of re-sampling. It also fitted to PLEC model successfully $p < 0.001$, but with slightly less robustness given that some failures out of the 100 times of re-sampling occurred, especially at the high diversity orders. This does not imply that PLEC model is not applicable to the archaea DAR at all; it only suggests that the order of accruing areas (hot spring samples) influences the fitting of the model.

(ii) The DAR profile: the parameter z-q series from the PL model is: z-q = {0.867[$q = 0$], 0.575[$q = 1$], 0.512[$q = 2$], 0.492[$q = 3$]}, a monotonically decreasing trend with the increase of diversity order (q). This indicates that at lower diversity orders, the spatial heterogeneity is larger than that at higher orders. The highest heterogeneity is at the species richness level ($q = 0$). Since at higher diversity order, the dominant species were given more weights in computing the diversity (Hill numbers). This suggests that the hot spring microbiomes are more homogenous (less heterogeneous or less diverse) in terms of their dominant species. Since at lower diversity order, the rare species were given more weights in computing the diversity (Hill numbers). This suggests that the hot spring microbiomes are more heterogeneous (more diverse) in terms of rare species.

(iii) The PDO profile: the parameter g-q series from the PL model is: g-q = {0.170[$q = 0$], 0.505[$q = 1$], 0.568[$q = 2$], 0.588[$q = 3$]}, a monotonically increasing trend with the increase of diversity order (q). This indicates that at lower diversity orders, the pair-wise diversity overlap (similarity) is smaller than that at higher orders. The lowest diversity overlap (similarity) is at the species richness level ($q = 0$), and the highest similarity occurred at the diversity order $q = 3$. This finding further supports the finding revealed by the DAR-profile in the previous item (ii).

(iv) The MAD profile: the parameter D_{max}-q series estimated from the PLEC model is: D_{max}-q = {8,397[$q = 0$], 248.3[$q = 1$], 68.8[$q = 2$], 45.0[$q = 3$]}, shows a monotonically decreasing trend with the increase of diversity order (q). This, of course, is determined by the nature of the diversity (Hill numbers). The $D_{max}(q = 0) = 8{,}397$ is simply the maximal accrual of species richness since the Hill numbers at $q = 0$ is the species richness. For example, at diversity order $q = 1$, the Hill numbers (our measure for the diversity) are equivalent to (function of) Shannon diversity index. This suggests that the maximal accrual diversity in terms of the Shannon index is 248.3. Similar interpretations can be made for $q = 2$, and 3. The maximal accrual species richness of 8,397 means that, it needs to accrue $A_{\max} = 262$ hot spring sites (areas) globally to reach this theoretical asymptote of species richness in the hot spring microbiome.

(v) There appears a trend of decreasing correlation coefficient (R) with increasing diversity order (q). This should be expected because with increasing q, the complexity associated with non-linearity in higher order entropy (i.e., Hill numbers) is increased. Consequently, goodness-of-fitting to the linear models (Equations 5, 6) is likely to decline.

DAR Analysis of the Bacteria

We did the same DAR analysis with bacteria dataset, and the results are exhibited in **Table 2**. We further performed Wilcox non-parametric significance test of the differences between Archaea and Bacteria in their DAR parameters, and it turned out that (i) regarding the PL-model, archaea-DAR and bacteria-DAR have significantly different DAR parameter values except for the diversity order $q = 1$. (ii) Regarding the PLEC model, archaea-DAR, and bacteria-DAR have significantly different DAR parameter values at the higher diversity orders ($q = 2$, 3), but no significant differences occurred at the lower diversity orders ($q = 0$, 1). These test results justify our attempt to separately build DAR models for archaea and bacteria. Since the format of **Tables 1**, **2** are exactly the same, our explanations for the bacteria-DAR is presented relatively brief intentionally. From **Table 2**, we summarized the following findings:

(i) The bacteria diversity scaling (changes across space) fitted to both the PL and PLEC models successfully, with all $p < 0.001$. The goodness-of-fitting is equally well with that for the archaea DAR models in the previous sub-section.

(ii) The DAR profile: the parameter z-q series from the PL model is: z-q = {0.830[$q = 0$], 0.616[$q = 1$], 0.575[$q = 2$], 0.544[$q = 3$]}, a monotonically decreasing trend with the increase of diversity order (q). This indicates that at lower diversity orders, the spatial heterogeneity is larger than that at higher orders. The highest heterogeneity is at the species richness level ($q = 0$). This pattern *per se* is the same as that of archaea-DAR.

If we further compare both the DAR profiles (see statistical tests in **Table 3**), we found that archaea has a larger diversity scaling parameter (z-value*s*) at diversity order $q = 0$ (i.e., species richness, but smaller scaling parameter (z-values) at diversity order $q = 2$ or 3. At diversity order $q = 1$, which is equivalent to the diversity measured with Shannon entropy and weighs all species in proportion with their relative abundance levels, archaea and bacteria showed no significant difference in their scaling parameter (z-values). These findings indicate that archaea is more heterogeneous or scaling faster than bacteria does in terms of species numbers (species richness), but is less heterogeneous or scaling slower than bacteria when the diversity (Hill numbers) were weighted computationally in favor of more abundant dominant species. When the diversity (Hill numbers) is weighted equally in terms of species abundances, archaea, and bacteria are equally heterogeneous over space or scaling at the same rate.

TABLE 1 | The parameters of DAR (diversity-area relationship) for the ***Archea*** in the hot springs, computed with 100 times of re-sampling from the totally permutated hot springs.

Diversity order and statistics		Power Law (PL)						PL with Exponential Cutoff (PLEC)							
		z	ln(c)	*R*	*p*-value	*g*	*N**	*z*	*d*	ln(c)	*R*	*p*-value	*N**	A_{max}	D_{max}
$q = 0$	Mean	0.867	4.657	0.983	<0.001	0.170	100	1.058	−0.004	4.206	0.990	<0.001	90	262	8396.9
	Std. Err.	0.012	0.054	0.001	<0.001	0.015		0.024	0.000	0.080	0.001	<0.001			
	Min	0.650	2.804	0.942	<0.001	−0.408		0.666	−0.014	1.368	0.958	<0.001			
	Max	1.268	5.690	0.997	<0.001	0.430		1.900	0.000	5.565	0.998	<0.001			
$q = 1$	Mean	0.575	2.771	0.923	<0.001	0.505	100	0.836	−0.006	2.201	0.951	<0.001	81	143	248.3
	Std. Err.	0.013	0.063	0.005	<0.001	0.014		0.025	0.000	0.082	0.003	<0.001			
	Min	0.234	0.882	0.799	<0.001	0.095		0.359	−0.017	0.115	0.847	<0.001			
	Max	0.930	4.278	0.987	<0.001	0.824		1.482	0.000	3.899	0.989	<0.001			
$q = 2$	Mean	0.512	1.740	0.900	<0.001	0.568	100	0.724	−0.005	1.293	0.931	<0.001	71	157	68.8
	Std. Err.	0.013	0.064	0.007	<0.001	0.013		0.025	0.000	0.084	0.005	<0.001			
	Min	0.172	−0.033	0.604	<0.001	0.185		0.317	−0.017	-0.302	0.713	<0.001			
	Max	0.860	3.193	0.980	<0.001	0.873		1.341	0.000	2.919	0.983	<0.001			
$q = 3$	Mean	0.492	1.392	0.893	<0.001	0.588	100	0.675	−0.004	1.042	0.924	<0.001	64	163	45.0
	Std. Err.	0.013	0.061	0.007	<0.001	0.013		0.026	0.000	0.085	0.007	<0.001			
	Min	0.152	−0.208	0.535	<0.001	0.240		0.244	−0.016	-0.454	0.652	<0.001			
	Max	0.815	2.805	0.980	<0.001	0.889		1.289	0.000	2.676	0.989	<0.001			

N = the number of successful fitting to DAR model from 100 times of random re-sampling of the individual orders. All the parameters were averaged from the N^* times of re-sampling, i.e., the parameters from successful fitting. Detailed parameters of the 100 DAR models from the 100 times of re-sampling are provided in the online Supplementary Tables (**Tables S1–S6**).*

TABLE 2 | The parameters of DAR (diversity-area relationship) for the ***Bacteria*** in the hot springs, computed with 100 times of re-sampling from the totally permutated hot springs.

Diversity Order and Statistics		**Power Law (PL)**						**PL with Exponential Cutoff (PLEC)**							
		z	**ln(c)**	***R***	***p*-value**	***g***	***N****	***z***	***d***	**ln(c)**	***R***	***p*-value**	***N****	A_{max}	D_{max}
q = 0	Mean	0.830	6.738	0.983	<0.001	0.217	100	1.031	−0.004	6.241	0.989	<0.001	88	256	55489.3
	Std. Err.	0.011	0.051	0.002	<0.001	0.014		0.022	0.000	0.073	0.001	<0.001			
	Min	0.568	5.634	0.930	<0.001	−0.091		0.674	−0.011	4.814	0.959	<0.001			
	Max	1.064	7.985	0.998	<0.001	0.517		1.444	0.000	7.517	0.999	<0.001			
q = 1	Mean	0.616	5.016	0.929	<0.001	0.456	100	0.914	−0.006	4.317	0.959	<0.001	87	144	2831.1
	Std. Err.	0.018	0.084	0.005	<0.001	0.020		0.033	0.000	0.115	0.003	<0.001			
	Min	0.262	2.907	0.691	<0.001	−0.104		0.389	−0.018	1.805	0.777	<0.001			
	Max	1.073	6.671	0.987	<0.001	0.801		1.607	0.000	6.301	0.997	<0.001			
q = 2	Mean	0.575	3.283	0.906	<0.001	0.502	100	0.827	−0.005	2.733	0.945	<0.001	73	152	427.7
	Std. Err.	0.015	0.075	0.007	<0.001	0.016		0.027	0.000	0.099	0.006	<0.001			
	Min	0.223	1.122	0.609	<0.001	0.019		0.398	−0.014	1.058	0.709	<0.001			
	Max	0.987	4.982	0.982	<0.001	0.833		1.391	0.000	4.566	0.997	<0.001			
q = 3	Mean	0.544	2.697	0.892	<0.001	0.535	100	0.777	−0.005	2.209	0.935	<0.001	70	152	207.9
	Std. Err.	0.014	0.067	0.008	<0.001	0.014		0.025	0.000	0.088	0.007	<0.001			
	Min	0.193	0.776	0.504	<0.001	0.132		0.410	−0.013	0.787	0.694	<0.001			
	Max	0.901	4.382	0.985	<0.001	0.857		1.288	0.000	3.847	0.991	<0.001			

N = the number of successful fitting to DAR model from 100 times of random re-sampling of the individual orders. All the parameters were averaged from the N* times of re-sampling, i.e., the parameters from successful fitting. Detailed parameters of the 100 DAR models from the 100 times of re-sampling are provided in the online Supplementary Tables (**Tables S1**–**S6**).*

The above finding also highlighted the necessity of using Hill numbers as general diversity measures (the diversity profile) over using a single *ad-hoc* diversity measure such as Shannon entropy or Simpson's index, because the latter may lead to inconsistent results or loss of information. This also shows the necessity of using DAR profile, a series of scaling parameter values (z) across different diversity orders (q), rather than using a single scaling parameter as in classic SAR (species-area relationship) analysis.

(iii) The PDO profile: the parameter g-q series from the PL model is: g-q = {0.217[$q = 0$], 0.456[$q = 1$], 0.502[$q = 2$], 0.535[$q = 3$]}, a monotonically increasing trend with the increase of diversity order (q). This pattern is the same as that of the archaea DAR PDO profile.

Similar to the previous DAR-profile comparison between archaea and bacteria, the archaea has a smaller PDO overlap (similarity) than the bacteria has at species richness level ($q = 0$), but has a larger PDO overlap (similarity) at the higher diversity order $q = 2$ or 3. At the diversity order $q = 1$, archaea and bacteria have the same level of diversity overlap (similarity) across space. The interpretation for this finding is exactly the same as that for the DAR-profile above.

(iv) The MAD profile: the parameter D_{max}-q series estimated from the PLEC model is: D_{max}-q = {55,489[$q = 0$], 2,831.1[$q = 1$], 427.7[$q = 2$], 207.9 [$q = 3$]}, showing a monotonically decreasing with the increase of diversity order (q). This, of course, is determined by the nature of the diversity (Hill numbers). The $D_{max}(q = 0) = 55{,}489$ is simply the maximal accrual of species richness since the Hill numbers at $q = 0$ is the species richness. Similar interpretations can be made for $q = 1$, 2, and 3. The maximal accrual species richness of 55,489 means that, it needs to accrue $A_{\max} = 256$ hot spring sites (areas) globally to reach this theoretical asymptote of species richness in the hot spring microbiome.

(v) Similar to the previous DAR models for archaea, there appears a trend of decreasing correlation coefficient (R) with increasing diversity order (q). This should be expected as explained previously.

While the above *pattern* for bacteria MAD-profile is similar to the pattern for archaea MAD profile, the vis-a-vis comparison of both the MAD profiles is a different story. Obviously, the values of bacteria D_{max} are far larger than the values of archaea D_{max}. Indeed, the difference is consistent with biological (ecological) reality that there are far more bacteria species than archaea species in hot springs. Unfortunately, unlike the cases of DAR and PDO profiles, we cannot perform the permutation (randomization) test for MAD-profile (D_{max}). This is because the D_{max} was computed based on the average parameter values form 100 times of re-sampling. We believe biological (ecological) observations justify our claims that MAD-profiles are also different between the archaea and bacteria.

DAR Analysis With the Combined Datasets of Archaea and Bacteria

Since there are significant differences between the archaea and bacteria in their DAR parameters, ideally, independent DAR models should be built for each kingdom. However, there is no doubt that they cohabitate (coexist) in the hot spring environment. Therefore, building unified DAR models (**Table S7**) for the combined archaea and bacteria is justified. **Table S7** shows that the DAR models fitted to the combined datasets of archaea and bacteria equally well with those for the archaea or bacteria, independently. For practical purpose such as conservation planning, the unified models (**Table S7**) are obviously more convenient, but for theoretical (mechanistic) inquiries, the separately built DAR models previously (**Tables 1**, **2**) should be more appropriate. Since the pattern of the unified DAR models are similar to the separately built ones, except some nuances, which make little differences for practical applications. As to the theoretical implications of those nuances, we recommend the use of those separately built DAR models directly. Therefore, we do not further compare the subtle differences between the unified and separate DAR models here.

TABLE 3 | The p-values of Wilcox non-parametric significance test between the differences between Archaea and Bacteria in their DAR parameters.

Diversity order (q)	DAR-PL model						DAR-PLEC model		
	z			g			z		
	≠	>	<	≠	>	<	≠	>	<
$q = 0$	0.030	0.015	0.985	0.030	0.985	0.015	0.245	0.123	0.878
$q = 1$	0.137	0.932	0.069	0.137	0.069	0.932	0.114	0.943	0.057
$q = 2$	0.005	0.997	0.003	0.005	0.003	0.997	0.032	0.984	0.016
$q = 3$	0.011	0.995	0.005	0.011	0.005	0.995	0.028	0.986	0.014

TABLE 4 | The LGD (the ratio of local diversity to global maximal accrual diversity) profile for the archaea, bacteria, and combined communities in the hot springs.

Diversity order (q)	Alpha-LGD (%) (Archaea)	Alpha-LGD (%) (Bacteria)	Alpha-LGD (%) (Combined)
$q = 0$	1.25	1.52	1.57
$q = 1$	6.43	5.33	5.75
$q = 2$	8.29	6.23	5.95
$q = 3$	8.95	7.13	6.12

The Ratio of Local Diversity to Regional (or Global) Accrual Diversity

The LRD (or LGD) is defined as the ratio of the local diversity of an *averaged area* to the regional diversity [or the global maximal accrual diversity (MAD)]. The dividend (local diversity) can be estimated with parameter c of the DAR-PL model, and the divisor can be either regional accrual diversity (which can be estimated with PLEC model directly) or global maximal accrual diversity (which is simply the MAD or D_{max}). Note that we defined LRD/LDG at different diversity orders ($q = 0, 1, 2, 3, \ldots$) as LRD/LGD profile.

Here we only computed LGD, the global version of the ratio. **Table 4** listed the LGD for the archaea, bacteria and their combined microbiome, at each diversity order (q). For example,

at species richness level ($q = 0$), the LGD is between 1.24 and 1.57%, which suggests that, on average, a single (local) hot spring only hosts between 1.24 and 1.57% of the global scale diversity. At high diversity orders, the ratios increased (up to 9% approximately). Another interesting observation is that at the species richness ($q = 0$), the LGD for archaea is lower than that for bacteria. However, at higher diversity orders ($q = 1, 2, 3$), the trend is reversed.

DISCUSSION

With the gold rush of microbial community ecology, thanks to the revolutionary metagenomic sequencing technology, the classic SAR has been called for new missions. Green et al. (2004) and Horner-Devin et al. (2004) published, in the same issue of the journal *Nature*, the first two studies on the SAR of microbes. The following year, two other important studies by Bell et al. (2005) and Smith et al. (2005) were published in two other leading journals, *Science* and *PNAS*, respectively. The SAR power law exponent (b) values from those studies were 0.074 (fungi), 0.019–0.040 (bacteria in marsh sediment), 0.26 (bacteria in tree holes), and 0.134 (phytoplankton). According to Green and Bohannan (2006) review, the reported SAR exponents in microbes were in the range between 0.019 and 0.470, but most values were below 0.2 (8 out of 11 studies). A major limitation of these pioneering studies on the testing of SAR with microbes is then low throughput of DNA sequencing technology in detecting bacteria, and consequently the diversity and SAR exponent were significantly underestimated. Even with the technology limitation, the reported exponent values have already indicated the applicability of SAR in microbes, and recent studies further confirmed the validity of microbial SAR (e.g., Noguez et al., 2005; Peay et al., 2007; Bell, 2010; Barreto et al., 2014; Pop Ristova et al., 2014; Ruff et al., 2015; Terrat et al., 2015; Várbíró et al., 2017). For example, nearly a decade after Green and Bohannan (2006) review, the range of exponent (z) of microbial SAR is nearly unchanged and most studies have still been limited to bacteria and archaea (Barreto et al., 2014).

While the classic SAR has been well-recognized as one of the most significant laws in ecology and biogeography, it is not without limitations. The recent extension from SAR to DAR by Ma (2018a,b) generalized the scaling law of biodiversity from species richness (the number of species) to general diversity measures (the Hill numbers). Furthermore, DAR profile, PDO profile, MAD profile, LRD/LGD profile based on DAR models can offer useful novel tools to sketch out the biogeography maps, which comprehensively characterize the biodiversity scaling over space and time.

Despite the large number studies of microbial SAR in various environments, to the best of our knowledge, the SAR of the hot-spring microbiome has not been reported in existing literature. In consideration of the more general nature of DAR over SAR, we skipped SAR and directly applied DAR modeling to reanalyze the hot spring microbiome datasets of Sharp et al. (2014). In fact, our DAR analysis, as presented in previous sections, included SAR as a special case when diversity order $q = 0$. Our study therefore provides the first glimpse of the SAR/DAR of the hot spring microbiome. The results and conclusions we obtained should certainly be further verified in future with more extensive datasets of the hot spring microbiomes. Although the sample size of 165 hot springs, we used, is not small, the future studies should attempt to collect samples from more diverse regions from different continents to validate our study on a truly global scale.

AUTHOR CONTRIBUTIONS

ZM designed the study and wrote the paper. LL performed the data analysis. All authors approved the submission.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00118/full#supplementary-material

REFERENCES

Amin, A., Ahmed, I., Salam, N., Kim, B. Y., Singh, D., Zhi, X. Y., et al. (2017). Diversity and distribution of thermophilic bacteria in hot springs of Pakistan. *Microb. Ecol.* 74, 116–127. doi: 10.1007/s00248-017-0930-1

Arrhenius, O. (1921). Species and area. *J. Ecol.* 9, 95–99. doi: 10.2307/2255763

Barreto, D. P., Conrad, R., Klose, M., Claus, P., and Enrich-Prast, A. (2014). Distance-decay and taxa-area relationships for bacteria, archaea and methanogenic archaea in a tropical lake sediment. *PLoS ONE* 9:e110128. doi: 10.1371/journal.pone.0110128

Bell, T. (2010). Experimental tests of the bacterial distance-decay relationship. *ISME J.* 4, 1357–1365. doi: 10.1038/ismej.2010.77

Bell, T., Ager, D., Song, J., Newman, J. A., Thompson, I. P., Lilley, A. K., et al. (2005). Larger islands house more bacterial taxa. *Science* 308:1884. doi: 10.1126/science.1111318

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, 335–336. doi: 10.1038/nmeth.f.303

Chan, C. S., Chan, K. G., Ee, R., Hong, K. W., Urbieta, M. S., Donati, E. R., et al. (2017). Effects of physiochemical factors on prokaryotic biodiversity in Malaysian circumneutral hot springs. *Front. Microbiol.* 8:1252. doi: 10.3389/fmicb.2017.01252

Chao, A., Chiu, C. H., and Hsieh, T. C. (2012). Proposing a resolution to debates on diversity partitioning. *Ecology* 93, 2037–2051. doi: 10.1890/11-1817.1

Chao, A., Chiu, C. H., and Jost, L. (2014). Unifying species diversity, phylogenetic diversity, functional diversity and related similarity and differentiation measures through Hill numbers. *Annu. Rev. Ecol. Evol. Syst.* 45, 297–324. doi: 10.1146/annurev-ecolsys-120213-091540

Cole, J. K., Peacock, J. P., Dodsworth, J. A., Williams, A. J., Thompson, D. B., Dong, H., et al. (2013). Sediment microbial communities in Great Boiling Spring are controlled by temperature and distinct from water communities. *ISME J.* 7, 718–729. doi: 10.1038/ismej.2012.157

De León, K. B., Gerlach, R., Peyton, B. M., and Fields, M. W. (2013). Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geyser Basin, Yellowstone National Park. *Front. Microbiol.* 4:330. doi: 10.3389/fmicb.2013.00330

de Oliveira, T. B., Gomes, E., and Rodrigues, A. (2015). Thermophilic fungi in the new age of fungal taxonomy. *Extremophiles* 19, 31–37. doi: 10.1007/s00792-014-0707-0

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., et al. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 21390–21395. doi: 10.1073/pnas.1215210110

Gajigan, A. P., Yñiguez, A. T., Villanoy, C. L., San Diego-McGlone, M. L., Jacinto, G. S., and Conaco, C. (2018). Diversity and community structure of marine microbes around the Benham Rise underwater plateau, northeastern Philippines. *Peer J.* 6:e4781. doi: 10.7717/peerj.4781

Geesey, G. G., Barkay, T., and King, S. (2016). Microbes in mercury-enriched geothermal springs in western North America. *Sci. Tot. Environ.* 569–570, 321–331. doi: 10.1016/j.scitotenv.2016.06.080

Ghilamicael, A. M., Budambula, N. L. M., Anami, S. E., Mehari, T., and Boga, H. I. (2017). Evaluation of prokaryotic diversity of five hot springs in Eritrea. *BMC Microbiol.* 17:203. doi: 10.1186/s12866-017-1113-4

Green, J., and Bohannan, B. J. (2006). Spatial scaling of microbial biodiversity. *Trends Ecol. Evol.* 21, 501–507. doi: 10.1016/j.tree.2006.06.012

Green, J. L., Holmes, A. J., Westoby, M., Oliver, I., Briscoe, D., Dangerfield, M., et al. (2004). Spatial scaling of microbial eukaryote diversity. *Nature* 432, 747–750. doi: 10.1038/nature03034

Hartmann, M., Niklaus, P. A., Zimmermann, S., Schmutz, S., Kremer, J., Abarenkov, K., et al. (2014). Resistance and resilience of the forest soil microbiome to logging-associated compaction. *ISME J.* 8, 226–244. doi: 10.1038/ismej.2013.141

Hedlund, B. P., Murugapiran, S. K., Alba, T. W., Levy, A., Dodsworth, J. A., Goertz, G. B., et al. (2015). Uncultivated thermophiles: current status and spotlight on 'Aigarchaeota'. *Curr. Opin. Microbiol.* 25, 136–145. doi: 10.1016/j.mib.2015.06.008

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* 54, 427–432. doi: 10.2307/1934352

Horner-Devin, C., Lage, M. M., Hughes, J. B., and Bohannan, B. J. (2004). A taxa-area relationship for bacteria. *Nature* 432, 750–753. doi: 10.1038/nature03073

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

Inskeep, W. P., Jay, Z. J., Herrgard, M. J., Kozubal, M. A., Rusch, D. B., Tringe, S. G., et al. (2013a). Phylogenetic and functional analysis of metagenome sequence from high-temperature archaeal habitats demonstrate Linkages between metabolic potential and geochemistry. *Front. Microbiol.* 4:95. doi: 10.3389/fmicb.2013.00095

Inskeep, W. P., Jay, Z. J., Tringe, S. G., Herrgård, M. J., and Rusch, D. B. (2013b). The YNP metagenome project: environmental parameters responsible for microbial distribution in the Yellowstone geothermal ecosystem. *Front. Microbiol.* 4:67. doi: 10.3389/fmicb.2013.00067

Jiang, Z., Li, P., Jiang, D., Dai, X., Zhang, R., Wang, Y., et al. (2015). Microbial community structure and arsenic biogeochemistry in an acid vapor-formed spring in Tengchong Geothermal Area, China. *PLoS ONE* 11:e0146331. doi: 10.1371/journal.pone.0146331

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology* 88, 2427–2439. doi: 10.1890/06-1736.1

Liu, K. H., Ding, X. W., Salam, N., Zhang, B., Tang, X. F., Deng, B., et al. (2018). Unexpected fungal communities in the Rehai thermal springs of Tengchong influenced by abiotic factors. *Extremophiles* 22, 525–535. doi: 10.1007/s00792-018-1014-y

Ma, Z. S. (2017). Measuring microbiome diversity and similarity with hill numbers. *Metagenomics* 2018, 157–178 doi: 10.1016/B978-0-08-102268-9.00008-2

Ma, Z. S. (2018a). Extending species-area relationships (SAR) to diversity-area relationships (DAR). *Ecol. Evol.* 8:10023–10038.

Ma, Z. S. (2018b). Sketching the human microbiome biogeography with DAR (Diversity Area Relationship) profiles. *Microb. Ecol.*

Ma, Z. S., and Li, L. W. (2019). Semen microbiome biogeography: an analysis based on a Chinese population study. *Front. Microbiol.* (accepted). doi: 10.3389/fmicb.2018.03333

MacArthur, R. H., and Wilson, E. O. (1967). *The Theory of Island Biogeography*. Princeton, NJ: Princeton University Press.

Masaki, Y., Tsutsumi, K., Hirano, S. I., and Okibe, N. (2016). Microbial community profiling of the Chinoike Jigoku ("Blood Pond Hell") hot spring in Beppu, Japan: isolation and characterization of Fe(III)-reducing *Sulfolobus* sp. Strain GA1. *Res. Microbiol.* 167, 595–603. doi: 10.1016/j.resmic.2016.04.011

Merkel, A. Y., Pimenov, N. V., Rusanov, I. I., Slobodkin, A. I., Slobodkina, G. B., Tarnovetckii, I. Y., et al. (2017). Microbial diversity and autotrophic activity in Kamchatka hot springs. *Extremophiles* 21, 307–317. doi: 10.1007/s00792-016-0903-1

Meyer-Dombard, D. R., and Amend, J. P. (2014). Geochemistry and microbial ecology in alkaline hot springs of Ambitle Island, Papua New Guinea. *Extremophiles* 18, 763–778. doi: 10.1007/s00792-014-0657-6

Mohanrao, M. M., Singh, D. P., Kanika, K., Goyal, E., and Singh, A. K. (2016). Deciphering the microbial diversity of Tattapani water spring using metagenomic approach. *Int. J. Agri. Sci.* Res. 6, 371–382. doi: 10.1016/j.gdata.2015.04.014

Noguez, A. M., Escalante, A. E., García-Oliva, F., Arita, H. T., Forney, L., and Souza, V. (2005). Microbial macroecology: highly structured prokaryotic soil assemblages in a tropical deciduous forest. *Glob. Ecol. Biogeogr.* 14, 241–248. doi: 10.1111/j.1466-822X.2005.00156.x

Oliverio, A. M., Power, J. F., Washburne, A., Cary, S. C., Stott, M. B., and Fierer, N. (2018). The ecology and diversity of microbial eukaryotes in geothermal springs. *ISME J.* 12, 1918–1928. doi: 10.1038/s41396-018-0104-2

Peay, K. G., Bruns, T. D., Kennedy, P. G., Bergemann, S. E., and Garbelotto, M. (2007). A strong species–area relationship for eukaryotic soil microbes: island size matters for ectomycorrhizal fungi. *Ecol. Lett.* 10, 470–480. doi: 10.1111/j.1461-0248.2007.01035.x

Plotkin, J. B., Potts, M. D., Yu, D. W., Bunyavejchewin, S., Condit, R., Foster, R., Hubbell, S., et al. (2000). Predicting species diversity in tropical forests. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10850–10854. doi: 10.1073/pnas.97.20.10850

Poddar, A., and Das, S. K. (2018). Microbiological studies of hot springs in India: a review. *Arch. Microbiol.* 200, 1–18. doi: 10.1007/s00203-017-1429-3

Pop Ristova, P. P., Wenzhöfer, F., Ramette, A., Felden, J., and Boetius, A. (2014). Spatial scales of bacterial community diversity at cold seeps (Eastern Mediterranean Sea). *ISME J.* 9, 1306–1318. doi: 10.1038/ismej.2014.217

Preston, F. W. (1960). Time and space and the variation of species. *Ecology* 41, 611–627. doi: 10.2307/1931793

Preston, F. W. (1962). The canonical distribution of commonness and rarity: Part, I. *Ecology* 43, 185–215. doi: 10.2307/1931976

Ruff, S. E., Biddle, J. F., Teske, A. P., Knittel, K., Boetius, A., and Ramette, A. et al. (2015). Global dispersion and local diversification of the methane seep microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 112, 4015–4020. doi: 10.1073/pnas.1421865112

Salano, O. A., Makonde, H. M., Kasili, R. W., Wangai, L. N., Nawiri, M., and Boga, H. (2017). Diversity and distribution of fungal communities within the hot springs of soda lakes in the Kenyan rift valley. *Afr. J. Microbiol. Res.* 19, 764–777. doi: 10.5897/AJMR2017.8510

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. doi: 10.1128/AEM.01541-09

Sharp, C. E., Brady, A. L., Sharp, G. H., Grasby, S. E., Stott, M. B., and Dunfield, P. F. (2014). Humboldt's spa: microbial diversity is controlled by temperature in geothermal environments. *ISME J.* 8, 1166–1174. doi: 10.1038/ismej.2013.237

Sinha, R., Abnet, C. C., White, O., Knight, R., and Huttenhower, C. (2015). The microbiome quality control project: baseline study design and future directions. *Genome Biol.* 16:276. doi: 10.1186/s13059-015-0841-8

Smith, V. H., Foster, B. L., Grover, J. P., Holt, R. D., Leibold, M.A., and deNoyelles, F. (2005). Phytoplankton species richness scales consistently from laboratory microcosms to the world's oceans. *Proc. Natl. Acad. Sci. U.S.A.* 102, 4393–4396. doi: 10.1073/pnas.0500094102

Song, Z. Q., Wang, F. P., Zhi, X. Y., Chen, J. Q., Zhou, E. M., Liang, F., et al. (2013). Bacterial and archaeal diversities in Yunnan and Tibetan hot springs, China. *Environ. Microbiol.* 15, 1160–1175. doi: 10.1111/1462-2920.12025

Terrat, S., Dequiedt, S., Horrigue, W., Lelievre, M., Cruaud, C., Saby, N. P. A., et al. (2015). Improving soil bacterial taxa–area relationships assessment using DNA meta-barcoding. *Heredity* 114, 468–475. doi: 10.1038/hdy.2014.91

Tjørve, E. (2009). Shapes and functions of species–area curves (II): a review of new models and parameterizations. *J. Biogeogr.* 36, 1435–1445. doi: 10.1111/j.1365-2699.2009.02101.x

Ulrich, W., and Buszko, J. (2003). Self-similarity and the species–area relation of Polish butterflies. *Basic Appl. Ecol.* 4, 263–270. doi: 10.1078/1439-1791-00139

Urbieta, M. S., Donati, E. R., Chan, K., Shahar, S., Sin, L. L., and Goh, K. M. (2014a). Thermophiles in the genomic era: biodiversity, science, and applications. *Biotechnol. Adv.* 33, 633–647. doi: 10.1016/j.biotechadv.2015.04.007

Urbieta, M. S., Toril, E. G., Giaveno, M. A., Bazán, A. A., and Donati, E. R. (2014b). Archaea land bacterial diversity in five different hydrothermal ponds in the Copahue region in Argentina. *Syst. Appl. Microbiol.* 37, 429–441. doi: 10.1016/j.syapm.2014.05.012

Várbíró, G., Görgényi, J., Tóthmérész, B., Padisák, J., Hajnal, E., and Borics, G. (2017). Functional redundancy modifies species–area relationship for freshwater phytoplankton. *Ecol. Evol.* 7, 9905–9913. doi: 10.1002/ece3.3512

Watson, H. C. (1835). *Remarks on Geographic Distribution of British Plants*. London: Longman; Rees; Orme; Brown; Green and Longman.

Xie, W., Zhang, C. L., Wang, J., Chen, Y., Zhu, Y., de la Torre, J. R., et al. (2015). Distribution of ether lipids and composition of the archaeal community in terrestrial geothermal springs: impact of environmental variables. *Environ. Microbiol.* 17, 1600–1614. doi: 10.1111/1462-2920.12595

Conflict of Interest Statement: 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.

Copyright © 2019 Li and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 26 February 2019
doi: 10.3389/fmicb.2019.00291

Identification and Analysis of Human Microbe-Disease Associations by Matrix Decomposition and Label Propagation

Jia Qu, Yan Zhao and Jun Yin*

School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China

Studies have shown that microbes exist widely in the human body and are closely related to human complex diseases. Predicting potential associations between microbes and diseases is conducive to understanding the mechanisms of complex diseases and can also facilitate the diagnosis and prevention of human diseases. In this paper, we put forward the Matrix Decomposition and Label Propagation for Human Microbe-Disease Association prediction (MDLPHMDA) on the basis of the dataset of known microbe-disease associations collected from the database of HMDAD and the Gaussian interaction profile kernel similarity for diseases and microbes, disease symptom similarity. Moreover, the performance of our model was evaluated by means of leave-one-out cross validation and five-fold cross validation, and the corresponding AUCs of 0.9034 and 0.8954 ± 0.0030 were gained, respectively. In case studies, 10, 9, 9, and 8 out of the top 10 predicted microbes for asthma, colorectal carcinoma, liver cirrhosis, and type 1 diabetes were confirmed by literatures, respectively. Overall, evaluation results showed that MDLPHMDA has good performance in potential microbe-diseasepositive free parameter, which associations prediction.

Keywords: microbe, disease, association prediction, matrix decomposition, label propagation

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Xianwen Ren,
Peking University, China
Ping Xuan,
Heilongjiang University, China

****Correspondence:***
Yan Zhao
ts17060090a3@cumt.edu.cn

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

Received: *30 November 2018*
Accepted: *04 February 2019*
Published: *26 February 2019*

Citation:
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

INTRODUCTION

Microbes are microscopic organisms that may exist in single-celled form or in a colony of cells (Madigan and Michaelt, 2015). They live in almost all the habitats from the poles to the deep sea and also make up the microbiota in all multicellular organisms (Delong and Pace, 2001). There are trillions of microbes in the human body. Lots of them are beneficial for human health, while others may cause infectious diseases (Thiele et al., 2013). Human microbiota can form an endosymbiotic relationship with their host, providing services and useful goods to humans. For example, the gut flora can contribute to gut immunity as well as digest complex carbohydrates and synthesize vitamins (O'hara and Shanahan, 2006). It is now accepted that most of the microbes are not intrinsically harmful. However, the pathogenic microorganisms and the imbalance of resident microbes are closely related to human disease.

Microorganisms are closely related to both infectious diseases and non-infectious diseases. Infectious diseases are global problems. They have induced several feared plagues in human history and new infections are still emerging today (Morse, 1995). Microorganisms are the causative pathogens for many infectious diseases. The involved organisms include pathogenic bacteria such

as *Mycobacterium tuberculosis* and *Bacillus anthracis*, which can cause tuberculosis and anthrax, respectively (Hawn et al., 2014; Hendricks et al., 2014); protozoan parasites such as *Plasmodium* and *Toxoplasma gondii*, which can cause malaria and toxoplasmosis (Torgerson and Mastroiacovo, 2013; Iburg, 2015); and also fungi such as *Candida albicans* and *Histoplasma capsulatum* which can cause candidiasis or histoplasmosis (Stenn, 1960; Pappas et al., 2016). Meanwhile, most new infections appear to be caused by already discovered pathogenic microorganisms. These pathogens obtain selective advantage by changing conditions to infect new host populations or cause a new disease (Morse, 1991). On the other hand, microbiota can interact with human at multiple levels. Due to these complex microbiota-host relationships, dysbiosis can be the cause of the pathology (Forum on Microbial et al., 2014). Various factors including antibiotics, radiations, stress or nutritional changes can alter the compositions of human microbiota. This disruption of homeostasis can induce many maladies (Tamboli et al., 2004). For example, it is founded that the interactions between host immunity and gut microbiota can directly result in inflammatory bowel disease (IBD). IBD is a long-term aggravating inflammation of the intestine (Schirbel and Fiocchi, 2010). Both commensal microbiota and individual genetic susceptibility play key roles in the occurrence and development of this disease (Ferreira et al., 2014). Compared to healthy control, the composition of gut microbiota in IBD patients is distinct with decreased *Firmicutes* (Walker et al., 2011). The complex interplay between microbiota and human is also closely related to metabolic disease such as obesity (Ley et al., 2006). In a study about overweight and obese children, scientists found that the lower numbers of fecal *Staphylococcus aureus* was further linked with normal-weight development (Kalliomaki et al., 2008). Besides intestinal tract, microbial communities in respiratory tract are also closely related to various lung diseases such as sinusitis and chronic obstructive pulmonary disease (COPD) (Huang et al., 2017b). A study showed that sinusitis patients experienced an increase in *Corynebacterium tuberculostearicum* (Abreu et al., 2012). In COPD, increased *Lactobacillus* is induced by an inflammatory modulation and results in the formation of tertiary lymphoid (Sze et al., 2012). All the above studies revealed the close associations between microbes and various human diseases. Unquestionably, identifying potential microbe-disease associations is of great significance in exploring the pathogenesis, prevention, and treatment of diseases. As the traditional experimental method is time-consuming, costly, random, and blind, there is an urgent need to develop an effective calculation approach so as to help researchers in finding the regular pattern of microbe-disease associations and to provide complementary and supportive evidence for the experimental study.

Relevant research for the identification of potential microbe-disease associations are still in its infancy, and effective calculation models for the association prediction are even more scarce. Ma et al. (2017) created the first database of Human Microbe-Disease Association Database (HMDAD), which collected confirmed microbe-disease associations from published literatures. Based on the above work, several computational models were established to prioritize candidate microbes for diseases. For example, Chen et al. (2017a) introduced the network-based model of KATZ measure for Human Microbe-Disease Association prediction (KATZHMDA), the first calculation method for the identification of new microbe-disease associations through computing the number of walks of connections between microbe and disease nodes in the microbe-disease association network. Recently, the computational model of Laplacian Regularized Least Squares for Human Microbe-Disease Association (LRLSHMDA) was presented by Wang et al. (2017). It is a global measure based on a semi-supervised learning framework. In their proposed calculation model, the Laplacian regularized least squares (LapRLS) classification was adopted to prioritize candidate microbes for all interested diseases through the application of known microbe-disease associations, the Gaussian interaction profile kernel similarity for microbes and diseases. Similarly, with the same dataset of known microbe-disease associations, the Gaussian interaction profile kernel similarity for microbes and diseases mentioned above, a path-based search model of Path-Based Human Microbe-Disease Association prediction (PBHMDA) was introduced by Huang et al. (2017c). In the model, the association score of each microbe-disease pair would be computed by the integration of all paths less four between the microbe and disease with different weights. In addition, Huang et al. (2017a) put forward a Neighbor- and Graph-based combined Recommendation model for Human Microbe-Disease Association prediction (NGRHMDA). The final prediction scores of novel microbe-disease associations were attained via the integration of two prediction results predicted by neighbor-based collaborative filtering and the graph-based scoring method. Also, Peng et al. (2018) put forward a model of Adaptive Boosting for Human Microbe-Disease Association prediction (ABHMDA) by enforcing a strong classifier on the samples. Specifically, the strong classifier was constructed by the integration of 30 weak classifiers with different weights.

In this paper, by combining known microbe-disease associations collected from HMDAD, disease symptom similarity and Gaussian interaction profile kernel similarity for microbes and diseases, we introduced a computational model of Matrix Decomposition and Label Propagation for the Human Microbe-Disease Association prediction (MDLPHMDA). In our proposed algorithm, a new adjacency matrix of microbe-disease associations was first generated by employing the spare learning method (SLM) on the original association information extracted from HMDAD, and potential microbe-disease associations would be further predicted under the implementation of the label propagation algorithm (LPA). The leave-one-out cross validation (LOOCV) and five-fold cross validation were subsequently enforced for accuracy evaluation of MDLPHMDA. Assessment results showed that MDLPHMDA gained the area under the receiver operating characteristic curves (AUCs) of 0.9034 and 0.8954 $\pm$ 0.0030 in LOOCV and five-fold cross validation, respectively. In case studies, we carried out MDLPHMDA to predict potential microbes for asthma and colorectal carcinoma (CRC), respectively. Moreover, via the implementation of our developed algorithm, we prioritized microbes for liver cirrhosis and type 1 diabetes by removing their

known related microbes, respectively. Finally, the results analysis of cross validations and case studies showed that MDLPHMDA is a suitable and effective model in potential microbe-disease association prediction.

MATERIALS AND METHODS

Human Microbe-Disease Associations

The dataset of confirmed microbe-disease associations used in this paper were collected from HMDAD (http://www.cuilab.cn/hmdad) (Ma et al., 2017). According to the 16s RNA sequencing-based microbiome research, the database collected 483 microbe-disease associations between 39 diseases and 292 microbes from 61 previous works. Along with the deletion of the same microbe-disease associations based on different evidences in the database, we finally obtained a dataset of 450 associations between 39 diseases and 292 microbes. Moreover, the variables *nd* and *nm* were defined to represent the 39 diseases and 292 microbes, respectively. Also, adjacency matrix A of the verified microbe-disease associations was defined as follows:

$$A(i,j) = \begin{cases} 1, \textit{if microbe } m(j) \textit{ is related to disease } d(i) \\ 0, \qquad\qquad \textit{otherwise} \end{cases} \tag{1}$$

Integrated Diseases Similarity

The integrated disease similarity was constructed by combining the Gaussian interaction profile kernel similarity for diseases and disease symptom similarity. First, we calculated the Gaussian interaction profile kernel similarity for diseases by adopting the calculation approach in the previous literature (Van Laarhoven et al., 2011). According to the idea that similar diseases possess similar interaction and non-interaction patterns with microbes, the Gaussian interaction profile kernel similarity for diseases was created in light of confirmed microbe-disease associations. We defined the interaction profile of each disease by using a binary vector that shows whether the disease is related to each microbe or not. For example, for disease $d(i)$, its interaction profile $IP(d(i))$ is the ith row of the adjacency matrix A. Therefore, the Gaussian interaction profile kernel similarity between disease $d(i)$ and disease $d(j)$ can be computed as follows:

$$KD(d(i), d(j)) = \exp(-\gamma_d \left\| IP(d(i)) - IP(d(j)) \right\|^2) \tag{2}$$

$$\gamma_d = \gamma_d{}' / (\frac{1}{nd} \sum_{k=1}^{nd} \left\| IP(d(k)) \right\|^2) \tag{3}$$

where γ_d indicates the normalized kernel bandwidth in light of the new bandwidth parameter $\gamma_d{}'$. Second, according to the data of diseases and their symptoms in PubMed bibliography, disease symptom similarity DSS could be constructed (Zhou et al., 2014). Finally, in accordance with disease symptom similarity put forward by Zhou et al. (2014), taking into account of the Gaussian interaction profile kernel similarity for diseases, we constructed integrated disease similarity by using the method applied in a previous study (Chen et al., 2017a).

$$DS = \frac{KD + DSS}{2} \tag{4}$$

Gaussian Interaction Profile Kernel Similarity for Microbes

In the same way, motivated by previous literature (Van Laarhoven et al., 2011), the Gaussian interaction profile kernel similarity for microbes was established according to confirmed microbe-disease associations. For microbe $m(j)$, its interaction profile $IP(m(j))$ is the jth column of the adjacency matrix A. Therefore, the Gaussian interaction profile kernel similarity between microbe $m(i)$ and microbe $m(j)$ can be computed as follows:

$$KM(m(i), m(j)) = \exp(-\gamma_m \left\| IP(m(i)) - IP(m(j)) \right\|^2) \tag{5}$$

$$\gamma_m = \gamma_m{}' / (\frac{1}{nm} \sum_{k=1}^{nm} \left\| IP(m(k)) \right\|^2) \tag{6}$$

where γ_m indicates the normalized kernel bandwidth in light of the new bandwidth parameter $\gamma_m{}'$.

MDLPHMDA

In this manuscript, motivated by SLM developed by Pech et al. (2017) and LPA introduced by Zhang et al. (2017), we applied the calculation model of MDLPHMDA to infer novel microbe-disease associations. Starting from the fact that redundant formation may be present in the original dataset of known microbe-disease associations, we employed matrix decomposition to eliminate the noise of known microbe-disease associations and then applied LPA for the identification of the potential microbe-disease associations (see **Figure 1**). It is worth mentioning that matrix decomposition has been widely used in Bioinformatics research (Chen et al., 2018b,d; Zhao et al., 2018).

Since a part of microbe-disease associations in the dataset may be incorrect or redundant, we adopted SLM to remove the noise of the original data and search a lowest-rank matrix among candidates to gain a novel adjacency matrix. In our introduced model, we divided the original adjacency matrix A into two parts by using SLM. The first part is a linear combination of the original adjacency matrix A and a low-rank matrix, while the second part is a spare matrix that can be regarded as the noise of the original adjacency matrix A. Hence, the original adjacency matrix can be decomposed as follows:

$$A = AX + E \tag{7}$$

In order to get a low-rank matrix X and a sparse matrix E, we could transform Equation (7) into a optimization problem by applying the nuclear norm on X and the sparse norm on E.

$$\min_{X,E} \|X\|_* + \alpha \|E\|_{2,1} \; s.t. \; A = AX + E \tag{8}$$

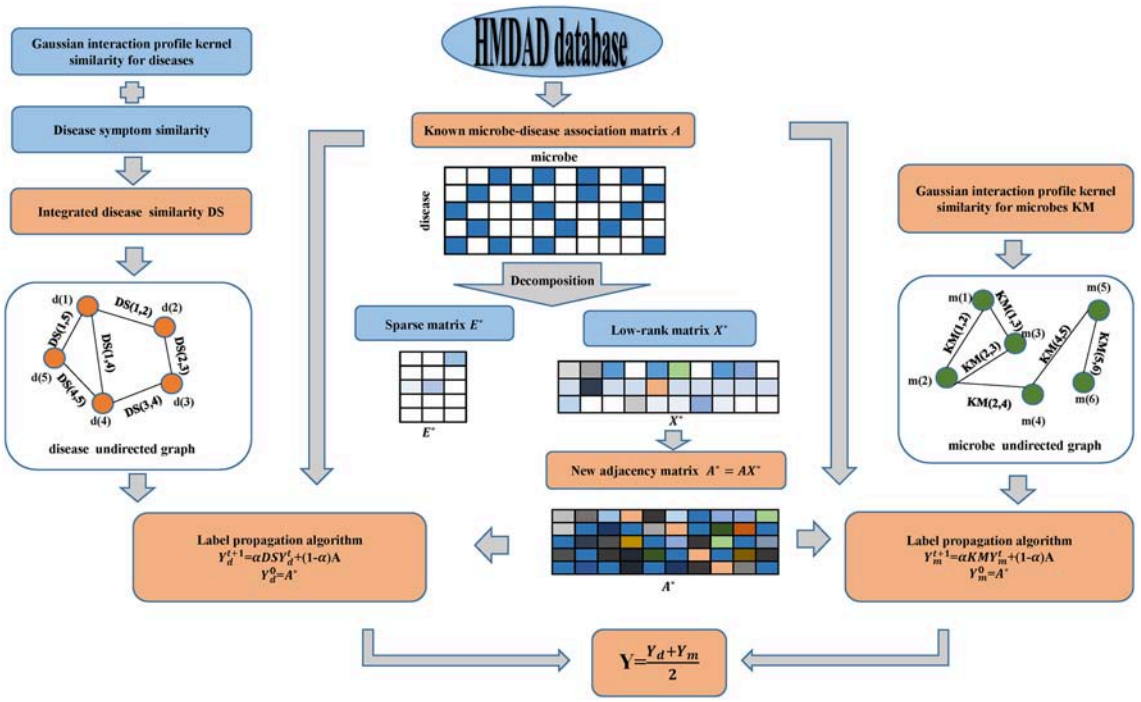


FIGURE 1 | Flowchart of the calculation model of MDLPHMDA: We first enforced matrix decomposition to eliminate the noise of original known microbe-disease associations and gained a new adjacency matrix. Then LPA was implemented based on the created new adjacency matrix for the identification of the potential microbe-disease associations.

where

$$\|A\|_* = \sum\nolimits_i \sigma_i \text{ (i.e., } \sigma_i \text{ is the sigular values of A)} \quad (9)$$

$$\|E\|_{2,1} = \sum\nolimits_{j=1}^{n} \sqrt{\sum\nolimits_{i=1}^{n} (E_{ij})^2} \quad (10)$$

Here, α is a positive free parameter, which can balance the weight between the low-rank matrix and the sparse matrix. To transform the original optimization problem into an augmented Lagrange function, we rewrote the optimization problem into a constraint and convex optimization problem of Equation (11) and enforced an inexact augmented Lagrange multipliers (IALM) algorithm (Meng et al., 2014) to solve it (see **Table 1**).

$$\min_{X,E,J} \|J\|_* + \alpha\|E\|_{2,1}$$
$$s.t.\ A = AX+E, X = J \quad (11)$$
$$L = \|J\|_* + \alpha\|E\|_{2,1} + tr(Y_1^T(A - AX - E))$$
$$+tr(Y_2^T(X - J)) + \frac{\mu}{2}(\|A - AX - E\|_F^2 + \|X - J\|_F^2) \quad (12)$$

where $\mu \geq 0$ is a penalty parameter and the detailed solution process to gain solution X^* and E^* of Equation (12) could be explained in previous literature (Pech et al., 2017).

As the solution of Equation (12) was solved, we gained a new adjacency matrix A^* with less noise by the linear combination of the original adjacency matrix A and the low-rank matrix X^* as follows:

$$A^* = AX^* \quad (13)$$

Then, based on the Gaussian interaction profile kernel similarity for microbes and diseases, disease symptom similarity and the newly created adjacency matrix A^*, we enforced LPA to infer novel microbe-disease associations. First, from the perspective

TABLE 1 | Computational procedures of the Inexact augmented Lagrange multipliers (IALM) algorithm.

Algorithm: IALM

Input: Given a adjacency matrix $\boldsymbol{A}$ and parameter α=0.1

Output: $\mathbf{X}^*$ and $\mathbf{E}^*$

Initialize: $\mathbf{X} = 0, \mathbf{E} = 0, \mathbf{Y_1} = 0, \mathbf{Y_2} = 0, \mu = 10^{-4}, \max_\mu = 10^{10}, \rho = 1.1, \varepsilon = 10^{-6}$

while $\|\mathbf{A} - \mathbf{AX} - \mathbf{E}\|_\infty \geq \varepsilon$ and $\|\mathbf{X} - \mathbf{J}\|_\infty \geq \varepsilon$ **do**

a. $\mathbf{J} = \arg\min \frac{1}{\mu}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - (\mathbf{X} + \mathbf{Y_2}/\mu)\|_{\mathrm{F}}^2$

b. $\mathbf{X} = (\mathbf{I} + \mathbf{A}^{\mathrm{T}}\mathbf{A})(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\mathbf{E} + \mathbf{J} + (\mathbf{A}^{\mathrm{T}}\mathrm{Y}_1 - \mathbf{Y}_2)/\mu)$

c. $\mathbf{E} = \arg\min \frac{\alpha}{\mu}\|\mathbf{E}\|_{2,1} + \frac{1}{2}\|\mathbf{E} - (\mathbf{A} - \mathbf{AX} + \mathbf{Y}_1/\mu)\|_{\mathrm{F}}^2$

d. $\mathbf{Y}_1 = \mathbf{Y}_1 + \mu(\mathbf{A} - \mathbf{AX} - \mathbf{E})$; $\mathbf{Y}_2 = \mathbf{Y}_2 + \mu(\mathbf{X} - \mathbf{J})$

e. $\mu = \min(\rho\mu, \max_\mu)$

end while

of disease, we constructed an undirected graph with diseases as nodes, and similarity scores as edge weight. To combine the original microbe-disease associations information, we treated the new adjacency matrix of microbe-disease associations as the labels to propagate in the disease undirected graph and each label is updated through the absorption of its neighborhoods' label information with a rate of α and going back to its original known microbe-disease association nodes with a rate of $1-\alpha$. Referring to previous literature (Yao et al., 2017; Zhang et al., 2017), we set α as 0.3. The label propagation process can be described as follows:

$$Y_d{}^{t+1} = \alpha DSY_d{}^t + (1-\alpha)\, A \tag{14}$$

where $Y_d{}^t$ indicates the predicted scores between microbes and diseases at step t. Specifically, Y_d^0 refers to the newly created adjacency matrixA^*. The iteration would be stable after some steps (the change in value between $Y_d{}^{t+1}$ and $Y_d{}^t$ measured by L_1 norm is <10e-6). The final value Y_d would be the predicted scores of new microbe-disease associations from the perspective of diseases.

Also, from the perspective of microbes, we can build another microbe undirected graph and employ LPA to gain another predicted scores Y_m of novel microbe-disease associations. Finally, we defined the final predicted scores Y for the potential microbe-disease associations by the average of the two predicted scores mentioned above.

$$Y = \frac{Y_d + Y_m}{2} \tag{15}$$

RESULTS

Performance Evaluation

In order to test the prediction performance of MDLPHMDA based on the 450 confirmed microbe-disease associations collected from HMDAD (Ma et al., 2017), our model was compared with two classic algorithms (LRLSHMDA and KATZHMDA) on the basis of the evaluation method of LOOCV and five-fold cross validation. In LOOCV, each confirmed microbe-disease association was taken as test sample by turn and the rest 449 identified associations were used to train. After executing MDLPHMDA, the score of the test sample would be ranked with the scores of candidate samples that were made up of all unconfirmed microbe-disease pairs. In five-fold cross validation, we first divided the 450 microbe-disease association pairs into five equal parts and later made each part as test sample in turn and the remaining four parts of associations as training samples. In the same way, each test sample's score would be ranked with the scores of all candidate samples that were composed of unconfirmed microbe-disease pairs. As the sample divisions may cause bias, we enforced five-fold cross validation 100 times to gain an average value as the final result. If the ranking of the test sample is higher than a given threshold, our model is considered to make a successful prediction. Then, according to varying thresholds, we plotted the receiver operating characteristics (ROC) curve by computing the ratio of true positive rate (TPR, sensitivity) to false positive rate (FPR, 1-specificity). Sensitivity denotes the percentage of test samples which obtained ranks higher than the set threshold. Meanwhile, specificity denotes the percentage of negative microbe-disease pairs with ranks lower than the threshold. Finally, to assess the performance of MDLPHMDA effectively, we computed corresponding AUCs. When AUC = 1, the model possesses perfect forecast ability; when AUC = 0.5, the model possesses random forecast ability. In LOOCV, assessment results showed that MDLPHMDA, LRLSHMDA, and KATZHMDA gained the AUCs of 0.9034, 0.8909, and 0.8382, respectively (see **Figure 2**). In five-fold cross validation, MDLPHMDA, LRLSHMDA, and KATZHMDA gained the AUCs of 0.8954 ± 0.0030, 0.8794 ± 0.0029, and 0.8301 ± 0.0033, respectively. Stated thus, it can be seen that our model possesses good prediction ability and

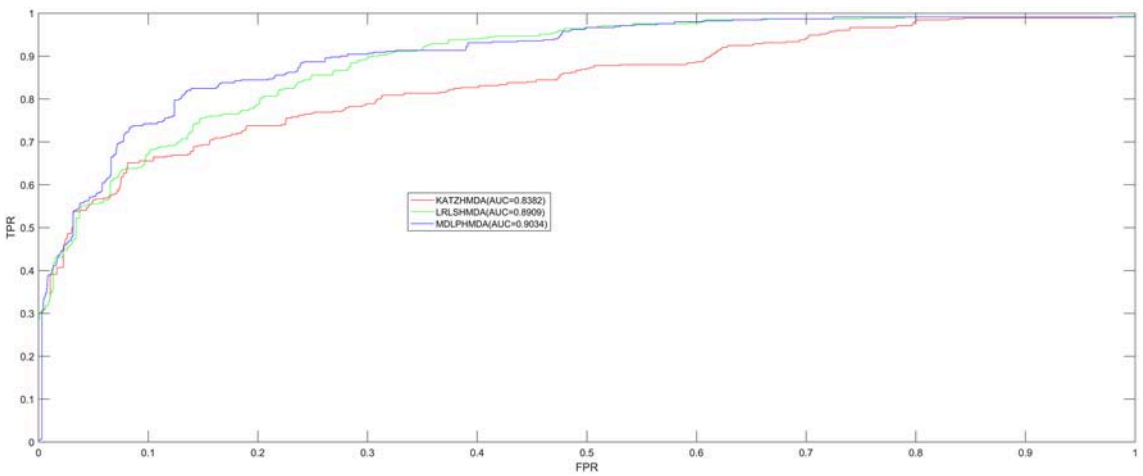


FIGURE 2 | Performance comparison between MDLPHMDA and other two classical microbe-disease association prediction models (LRLSHMDA and KATZHMDA) by means of AUCs based on LOOCV. The results showed that MDLPHMDA gained AUCs of 0.9034 in LOOCV.

could be used to assist the identification of novel microbe-disease associations. Moreover, we carried out a paired *t*-test based on the ranking results of LOOCV to observe the statistical significance of differences among MDLPHMDA, LRLSHMDA, and KATZHMDA. As a result, the *p*-value of MDLPHMDA and LRLSHMDA is 0.0088, whereas the *p*-value of MDLPHMDA and KATZHMDA is 1.2510e-08. We can see that MDLPMDA is significantly different from LRLSHMDA and KATZHMDA on the basis of their ranking results of LOOCV ($p < 0.05$).

Case Study

Via two different types of case studies, we further assessed the prediction ability of MDLPHMDA based on the confirmed 450 microbe-disease associations. In the first kind, we identified potential microbes for asthma and CRC, respectively, through the implementation of MDLPHMDA. Also, we released all prediction scores for 10938 novel microbe-disease pair between 39 diseases and 292 microbes (see **Supplementary Table 1**). In the second kind, we enforced MDLPHMDA to identify liver cirrhosis-associated microbes by removing 62 known liver cirrhosis-associated microbes from the dataset of known microbe-disease associations and also predicted for another disease of type 1 diabetes by removing its known microbes. Based on the results of the two types of case studies, the proposed algorithm of MDLPHMDA was proven to be an effective algorithm in the identification of novel microbe-disease associations.

Asthma is a long-term inflammatory disease of the airways (Lemanske and Busse, 2010). Its common symptoms include coughing, reversible airflow obstruction, wheezing, or bronchospasm (Lemanske and Busse, 2010). Epidemiological studies indicated that microbial exposures in early life might determine microbiota composition, which can help to prevent allergy or lead to the development of asthma (Wang et al., 2003; Weber et al., 2015). A study in asthmatic children has found a low abundance of *Bifidobacterium* in their intestinal microbiota, which may reduce the immune function and potentially contribute to disease chronicization (Kalliomaki et al., 2001). Similarly, a probiotic strain *Lactobacillus rhamnosus* reduced allergic responses in the airways of neonates (Martinon et al., 2009). In this paper, via the implementation of MDLPHMDA for the inference of novel asthma-related microbes, we could see that the top 10 predicted microbes for asthma were all confirmed through literature (see **Table 2**). Among the top 3 confirmed associations between microbes and asthma, relevant differences in *Firmicutes* were found between samples from asthmatic and non-asthmatic subjects (Marri et al., 2013). Another study investigated that *Clostridium difficile* was associated with an increased risk for asthma (Van Nimwegen et al., 2011). Meanwhile, in a study about early intestinal colonization of infants, *Clostridium coccoides* was confirmed to be associated with increased risk for the development of asthma before the age of 3 years (Vael et al., 2011).

CRC is the cancer in the colon or rectum (Watson and Collins, 2011). Common symptoms include weight loss, blood in stool, and feeling tired all the time (Watson and Collins, 2011). It typically starts in the form of a polyp as a benign tumor, which becomes cancerous over time (Watson and Collins, 2011). A quantitative polymerase chain reaction (qpcr) analysis verified that *Fusobacterium nucleatum*, an invasive anaerobe previously linked to appendicitis and periodontitis but not to cancer, was increased in a CRC tumor vs. normal tissue (Castellarin et al., 2012). Furthermore, this overabundance is positively associated with lymph node metastasis (Castellarin et al., 2012). Another study also observed a significant difference of *Bacteroides* and *Prevotella* in a CRC group, as compared to a normal group (Sobhani et al., 2011). Moreover, we employed the proposed algorithm to predict CRC-related microbes and the outcomes displayed that all but one of the top 10 microbes for CRC were verified (see **Table 3**). Among the top 3 confirmed associations, according to the taxonomic results, *Proteobacteria* showed a higher abundance in CRC rats compared to control groups and constitute the third most abundant phyla (Zhu et al., 2014). In another analysis on CRC, the *Helicobacter pylori* infection was noted in 50 CRC patients. Furthermore, an infection with *H. pylori CagA+* was associated with an increased risk for CRC (Shmuely et al., 2001). Moreover, a statistically significant difference in *C. difficile* was detected between the CRC and healthy group, suggesting a possible role of this bacteria in CRC carcinogenesis (Fukugaiti et al., 2015).

Liver cirrhosis is a disease induced by long-term damage. This damage is due to the replacement of normal tissue by scar tissue (Li et al., 1999). Typically, the disease develops slowly and there are often no significant early symptoms. As it worsens, patients may become tired, bruise easily, develop yellow skin, have fluid in the abdomen, or have swelling in the lower legs (Li et al., 1999). Liver cirrhosis is commonly caused by alcohol, non-alcoholic fatty liver disease, hepatitis B, or hepatitis C (Li et al., 1999). In a study on the alterations of the human microbiome in liver cirrhosis, quantitative metagenomics reveals 66 cognate bacterial species that differ in abundance between healthy individuals and patients, including *Alistipes finegoldii*, *Bacteroides eggerthii*, *Eubacterium rectale*, *Faecalibacterium prausnitzii*, *Haemophilus parainfluenzae,* and so on (Qin et al., 2014). In another study about fecal microbial communities in patients with liver

TABLE 2 | The validation of the top 10 predicted asthma-related microbes after implementing MDLPHMDA based on the confirmed microbe-disease associations from HMDAD.

Disease	Microbe	Score	Evidence
Asthma	*Firmicutes*	0.035113958	PMID:23265859
Asthma	*Clostridium difficile*	0.026401159	PMID:21872915
Asthma	*Clostridium coccoides*	0.023247917	PMID:21477358
Asthma	*Staphylococcus aureus*	0.022891159	PMID:25533526
Asthma	*Actinobacteria*	0.022703105	PMID:23265859
Asthma	*Lachnospiraceae*	0.022512904	PMID: 27433177
Asthma	*Lactobacillus*	0.022019371	PMID:20592920
Asthma	*Enterobacteriaceae*	0.018834777	PMID:21639872
Asthma	*Veillonella*	0.018243753	PMID: 26424567
Asthma	*Bacteroides*	0.017354957	PMID: 18822123

As a result, all of the top 10 predicted microbes were confirmed by literatures.

TABLE 3 | The validation of the top 10 predicted CRC-related microbes after implementing MDLPHMDA based on the confirmed microbe-disease associations from HMDAD.

Disease name	Microbe name	Score	Evidence
CRC	*Proteobacteria*	0.046859257	PMID: 24603888
CRC	*Helicobacter pylori*	0.023587271	PMID: 11774957
CRC	*Clostridium difficile*	0.023311807	PMID: 26691472
CRC	*Actinobacteria*	0.023310463	Unconfirmed
CRC	*Lactobacillus*	0.022992627	PMID:15828052
CRC	*Haemophilus*	0.022582968	PMID:26549775
CRC	*Lachnospiraceae*	0.02243978	PMID:21850056
CRC	*Clostridium coccoides*	0.021696923	PMID:18237311
CRC	*Enterobacteriaceae*	0.021122176	PMID: 25182170
CRC	*Staphylococcus aureus*	0.020450372	PMID:7074582

As a result, 9 out of the top 10 predicted microbes were confirmed by literatures.

TABLE 4 | The validation of the top 10 predicted liver cirrhosis-associated microbes after implementing MDLPHMDA by removing liver cirrhosis-related associations from the dataset of known microbe-disease associations.

Disease name	Microbe name	Score	Evidence
Liver cirrhosis	*Proteobacteria*	0.037652208	HMDAD
Liver cirrhosis	*Bacteroidetes*	0.033708121	HMDAD
Liver cirrhosis	*Firmicutes*	0.033302216	PMID:21574172
Liver cirrhosis	*Prevotella*	0.028842704	HMDAD
Liver cirrhosis	*Helicobacter pylori*	0.021108828	PMID:9365129
Liver cirrhosis	*Clostridium difficile*	0.020872569	PMID:26440041
Liver cirrhosis	*Actinobacteria*	0.020204542	PMID:22326468
Liver cirrhosis	*Clostridium coccoides*	0.018391455	Unconfirmed
Liver cirrhosis	*Staphylococcus aureus*	0.01835198	PMID:22833245
Liver cirrhosis	*Lactobacillus*	0.016449739	HMDAD

As a result, 9 out of the top 10 predicted microbes were confirmed by HMDAD and literatures.

cirrhosis, research has detected the prevalence of pathogenic bacteria such as *Enterobacteriaceae* and *Streptococcaceae* as well as the reduction of beneficial populations such as *Lachnospiraceae* (Chen et al., 2011). Here, by removing 62 known liver cirrhosis-associated microbes from the dataset of known microbe-disease associations, we enforced MDLPHMDA to identify liver cirrhosis-associated microbes on the basis of integrated disease similarity, Gaussian interaction profile kernel similarity for microbes, and the rest known microbe-disease associations. As a result, 9 out of the top 10 microbes for liver cirrhosis were confirmed by HMDAD and literature (see **Table 4**). Among the top 3 confirmed associations, *Firmicutes* was found to be highly enriched in the patients group (Chen et al., 2011). Moreover, researchers found significantly higher *H. pylori* prevalence in patients with previous hospital admissions (Siringo et al., 1997). This high prevalence of *H. pylori* is related to age and sex (Siringo et al., 1997). An analysis on the *C. difficile* infection in patients with liver cirrhosis showed that cirrhotic patients with the *C. difficile* infection have increased mortality than those without the *C. difficile* infection, suggesting the importance of *C. difficile* in the diagnosis and therapy of liver cirrhosis (Trifan et al., 2015).

Type 1 diabetes is a type of diabetes mellitus induced by very little or no insulin produced in the pancreas (Daneman, 2006). It results in high blood sugar levels in the human body. The classic symptoms include increased thirst and hunger, frequent urination and weight loss (Daneman, 2006). The cause of type 1 diabetes is still unclear. However, it is believed to involve both genetic and environmental factors (Chiang et al., 2014). One theory proposes that type 1 diabetes may be caused by an autoimmune response while the immune system attacks virus-infected insulin-producing cells in the pancreas (Knip et al., 2005). In a microbiome metagenomics analysis on type 1 diabetes, researchers identified the differences between patients and controls at the genus level. The most significant differences were noted in the genera *Prevotella* and *Bacteroides* (Brown et al., 2011). In another study defining the autoimmune microbiome for type 1 diabetes, scientists identified bacteria that correlated with the autoimmune state including *Bacteroides fragilis*, *Clostridia*, *Eubacterium eligens*, and so on (Giongo et al., 2011). Similarly, we employed MDLPHMDA to identify type 1 diabetes-associated microbes by removing 167 known type 1 diabetes-associated microbes from the dataset of known microbe-disease associations. The results showed that 8 out of the top 10 microbes for liver cirrhosis were confirmed (see **Table 5**). In a case-control study, scientists found a meaningful correlation between the *H. pylori* infection and the duration of diabetes in type 1 diabetic children (Bazmamoun et al., 2016). In another study, researchers found that *Staphylococcus aureus* is associated with the vitamin D receptor (VDR) polymorphisms in patients with type 1 diabetes (Panierakis et al., 2009).

DISCUSSION

Since the application of traditional experimental methods to prioritize disease-associated microbes is time consuming and expensive, the calculation approach of MDLPHMDA was put forward through the fusing of integrated disease similarity, Gaussian interaction profile kernel similarity for microbes and known microbe-disease associations. The performance of MDLPHMDA was tested using cross validations and case studies. Results on the basis of confirmed microbe-disease associations showed that the performance of our introduced algorithm is significantly improved in contrast with other two classic algorithms of LRLSHMDA and KATZHMDA. Consequently, the introduced algorithm is a suitable and effective model in the identification of novel microbe-disease associations. We further expect that the identified microbe-disease associations with high probability scores would be verified through biological experiment in the future.

The reason why MDLPHMDA could get excellent prediction performance is due to the following attractive properties. First, with the application of SLM on the original information of known microbe-disease associations, a new adjacency matrix with more accurate association information (the linear combination of low-rank matrix and the original adjacency matrix) and a noise (sparse) matrix would be gained. Obviously,

TABLE 5 | The validation of the top 10 predicted type 1 diabetes-related microbes after implementing MDLPHMDA by removing type 1 diabetes-related associations from the dataset of known microbe-disease associations.

Disease name	Microbe name	Score	Evidence
Type 1 diabetes	*Proteobacteria*	0.035314225	HMDAD
Type 1 diabetes	*Bacteroidetes*	0.030725459	HMDAD
Type 1 diabetes	*Firmicutes*	0.02813792	HMDAD
Type 1 diabetes	*Prevotella*	0.026244828	HMDAD
Type 1 diabetes	*Clostridium difficile*	0.021633674	Unconfirmed
Type 1 diabetes	*Helicobacter pylori*	0.021284419	PMID:27497772
Type 1 diabetes	*Clostridium coccoides*	0.019300705	Unconfirmed
Type 1 diabetes	*Staphylococcus aureus*	0.017744364	PMID:19411183
Type 1 diabetes	*Lactobacillus*	0.016082255	HMDAD
Type 1 diabetes	*Actinobacteria*	0.015923043	HMDAD

As a result, 8 out of the top 10 predicted microbes were confirmed by HMDAD and literatures.

in light of the new generated adjacency matrix, the forecast performance of the proposed algorithm for the identification of new microbe-disease associations could be significantly enhanced. Second, LPA was used to predict novel microbe-disease associations from the perspectives of microbe and disease, respectively, which would promote the ability of MDLPHMDA in terms of forecast accuracy. Third, in comparison with the previous calculation algorithms that only used Gaussian interaction profile kernel similarity for diseases as disease similarity, MDLPHMDA could achieve superior performance through integrating disease symptom similarity and Gaussian interaction profile kernel similarity for diseases into the final disease similarity. Moreover, the implementation of MDLPHMDA does not require negative samples and the algorithm could be applied to new diseases (microbes) without the relevant microbes (diseases).

However, the model has some main disadvantages. For instance, the amount of known microbe-disease associations used in this paper is very finite and more confirmed microbe-disease associations need to be collected. Additionally, as the computation of Gaussian interaction profile kernel similarity of microbes depended on known microbe-disease associations, other features of microbe similarity should be collected and combined to gain a more comprehensive dataset of microbe similarity such as microbe-drug associations collected by MDAD (Sun et al., 2018). For MDLPHMDA, it is difficult to find the optimum value of all the parameters to ensure that the prediction model achieves the highest accuracy. Also, the employment of SLM for creating new adjacency matrix may bring unnecessary and useless association information, which would affect the prediction result of LPA. Finally, successfully established models in the other computational fields would inspire the development of microbe-disease association prediction, such as microRNA-disease association prediction (Chen and Huang, 2017; Chen et al., 2018c), long non-coding RNA-disease association prediction (Chen and Yan, 2013; Chen et al., 2017b), drug-target interaction prediction (Chen et al., 2016b, 2018a), and synergistic drug combinations (Chen et al., 2016a).

AUTHOR CONTRIBUTIONS

JQ developed the prediction method, implemented the experiments, analyzed the result, and wrote the paper. YZ conceived the project, designed the experiments, analyzed the result, and revised the paper. JY analyzed the result and revised the paper.

FUNDING

JQ, YZ, and JY was supported by the National Natural Science Foundation of China under Grant No. 61772531.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00291/full#supplementary-material

REFERENCES

Abreu, N. A., Nagalingam, N. A., Song, Y., Roediger, F. C., Pletcher, S. D., Goldberg, A. N., et al. (2012). Sinus microbiome diversity depletion and Corynebacterium tuberculostearicum enrichment mediates rhinosinusitis. *Sci. Transl. Med.* 4:151ra124. doi: 10.1126/scitranslmed.3003783

Bazmamoun, H., Rafeey, M., Nikpouri, M., and Ghergherehchi, R. (2016). *Helicobacter pylori* infection in children with type 1 diabetes mellitus: a case-control study. *J. Res. Health Sci.* 16, 68–71.

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

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, X., Guan, N. N., Sun, Y. Z., Li, J. Q., and Qu, J. (2018a). MicroRNA-small molecule association identification: from experimental results to computational models. *Brief. Bioinformatics*. doi: 10.1093/bib/bby098. [Epub ahead of print].

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for miRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2017a). 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

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016a). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput. Biol.* 12:e1004975. doi: 10.1371/journal.pcbi.1004975

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018b). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018c). BNPMDA: bipartite network projection for MiRNA-Disease Association prediction. *Bioinformatics*. 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017b). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinformatics* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016b). Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinformatics* 17, 696–712. doi: 10.1093/bib/bbv066

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chen, X., Yin, J., Qu, J., and Huang, L. (2018d). MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, Y., Yang, F., Lu, H., Wang, B., Chen, Y., Lei, D., et al. (2011). Characterization of fecal microbial communities in patients with liver cirrhosis. *Hepatology* 54, 562–572. doi: 10.1002/hep.24423

Chiang, J. L., Kirkman, M. S., Laffel, L. M., and Peters, A. L. (2014). Type 1 diabetes through the life span: a position statement of the American Diabetes Association. *Diabetes Care* 37, 2034–2054. doi: 10.2337/dc14-1140

Daneman, D. (2006). Type 1 diabetes. *Lancet* 367, 847–858. doi: 10.1016/S0140-6736(06)68341-4

Delong, E. F., and Pace, N. R. (2001). Environmental diversity of bacteria and archaea. *Syst. Biol.* 50, 470–478. doi: 10.1080/10635150118513

Ferreira, C. M., Vieira, A. T., Vinolo, M. A., Oliveira, F. A., Curi, R., and Martins Fdos, S. (2014). The central role of the gut microbiota in chronic inflammatory diseases. *J. Immunol. Res.* 2014:689492. doi: 10.1155/2014/689492

Forum on Microbial, H., Board on Global, H., and Institute Of, M. (2014). "The National Academies Collection: Reports funded by National Institutes of Health," in *Microbial Ecology in States of Health and Disease: Workshop Summary* (Washington, DC: National Academies Press).

Fukugaiti, M. H., Ignacio, A., Fernandes, M. R., Ribeiro Junior, U., Nakano, V., and Avila-Campos, M. J. (2015). High occurrence of Fusobacterium nucleatum and *Clostridium difficile* in the intestinal microbiota of colorectal carcinoma patients. *Braz. J. Microbiol.* 46, 1135–1140. doi: 10.1590/S1517-838246420140665

Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., Novelo, L. L., Casella, G., et al. (2011). Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J.* 5, 82–91. doi: 10.1038/ismej.2010.92

Hawn, T. R., Day, T. A., Scriba, T. J., Hatherill, M., Hanekom, W. A., Evans, T. G., et al. (2014). Tuberculosis vaccines and prevention of infection. *Microbiol. Mol. Biol. Rev.* 78, 650–671. doi: 10.1128/MMBR.00021-14

Hendricks, K. A., Wright, M. E., Shadomy, S. V., Bradley, J. S., Morrow, M. G., Pavia, A. T., et al. (2014). Centers for disease control and prevention expert panel meetings on prevention and treatment of anthrax in adults. *Emerging Infect. Dis.* 20:e130678. doi: 10.3201/eid2002.130687

Huang, Y. A., You, Z. H., Chen, X., Huang, Z. A., Zhang, S., and Yan, G. Y. (2017a). 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

Huang, Y. J., Erb-Downward, J. R., Dickson, R. P., Curtis, J. L., Huffnagle, G. B., and Han, M. K. (2017b). Understanding the role of the microbiome in chronic obstructive pulmonary disease: principles, challenges, and future directions. *Transl. Res.* 179, 71–83. doi: 10.1016/j.trsl.2016.06.007

Huang, Z. A., Chen, X., Zhu, Z., Liu, H., Yan, G. Y., You, Z. H., et al. (2017c). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Iburg, K. M. (2015). Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 385, 117–171. doi: 10.1016/S0140-6736(14)61682-2

Kalliomaki, M., Collado, M. C., Salminen, S., and Isolauri, E. (2008). Early differences in fecal microbiota composition in children may predict overweight. *Am. J. Clin. Nutr.* 87, 534–538. doi: 10.1093/ajcn/87.3.534

Kalliomaki, M., Kirjavainen, P., Eerola, E., Kero, P., Salminen, S., and Isolauri, E. (2001). Distinct patterns of neonatal gut microflora in infants in whom atopy was and was not developing. *J. Allergy Clin. Immunol.* 107, 129–134. doi: 10.1067/mai.2001.111237

Knip, M., Veijola, R., Virtanen, S. M., Hyoty, H., Vaarala, O., and Akerblom, H. K. (2005). Environmental triggers and determinants of type 1 diabetes. *Diabetes* 54(Suppl. 2), S125–S136. doi: 10.2337/diabetes.54.suppl_2.S125

Lemanske, R. F. Jr., and Busse, W. W. (2010). Asthma: clinical expression and molecular mechanisms. *J. Allergy Clin. Immunol.* 125, S95–102. doi: 10.1016/j.jaci.2009.10.047

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, C. P., Lee, F. Y., Hwang, S. J., Chang, F. Y., Lin, H. C., Lu, R. H., et al. (1999). Spider angiomas in patients with liver cirrhosis: role of alcoholism and impaired liver function. *Scand. J. Gastroenterol.* 34, 520–523. doi: 10.1080/003655299750026272

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe-disease associations. *Brief. Bioinformatics* 18, 85–97. doi: 10.1093/bib/bbw005

Madigan and Michaelt (2015). *Brock Biology of Microorganisms, 14th Edn.* Pearson. Carbondale, IL.

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.e341–343. doi: 10.1016/j.jaci.2012.11.013

Martinon, F., Mayor, A., and Tschopp, J. (2009). The inflammasomes: guardians of the body. *Annu. Rev. Immunol.* 27, 229–265. doi: 10.1146/annurev.immunol.021908.132715

Meng, F., Yang, X., and Zhou, C. (2014). The augmented lagrange multipliers method for matrix completion from corrupted samplings with application to mixed Gaussian-impulse noise removal. *PLoS ONE* 9:e108125. doi: 10.1371/journal.pone.0108125

Morse, S. S. (1991). Emerging viruses: defining the rules for viral traffic. *Perspect. Biol. Med.* 34, 387–409. doi: 10.1353/pbm.1991.0038

Morse, S. S. (1995). Factors in the emergence of infectious diseases. *Emerg. Infect. Dis.* 1, 7–15. doi: 10.3201/eid0101.950102

O'hara, A. M., and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Rep.* 7, 688–693. doi: 10.1038/sj.embor.7400731

Panierakis, C., Goulielmos, G., Mamoulakis, D., Maraki, S., Papavasiliou, E., and Galanakis, E. (2009). Staphylococcus aureus nasal carriage might be associated with vitamin D receptor polymorphisms in type 1 diabetes. *Int. J. Infect. Dis.* 13, e437–e443. doi: 10.1016/j.ijid.2009.02.012

Pappas, P. G., Kauffman, C. A., Andes, D. R., Clancy, C. J., Marr, K. A., Ostrosky-Zeichner, L., et al. (2016). Executive summary: clinical practice guideline for the management of candidiasis: 2016 update by the Infectious Diseases Society of America. *Clin. Infect. Dis.* 62, 409–417. doi: 10.1093/cid/civ1194

Pech, R., Hao, D., Po, M., and Zhou, T. (2017). Predicting drug-target interactions via sparse learning.

Peng, L. H., Yin, J., Zhou, L., Liu, M. X., and Zhao, Y. (2018). Human microbe-disease association prediction based on adaptive boosting. *Front. Microbiol.* 9:2440. doi: 10.3389/fmicb.2018.02440

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

Schirbel, A., and Fiocchi, C. (2010). Inflammatory bowel disease: established and evolving considerations on its etiopathogenesis and therapy. *J. Dig. Dis.* 11, 266–276. doi: 10.1111/j.1751-2980.2010.00449.x

Shmuely, H., Passaro, D., Figer, A., Niv, Y., Pitlik, S., Samra, Z., et al. (2001). Relationship between *Helicobacter pylori* CagA status and colorectal cancer. *Am. J. Gastroenterol.* 96, 3406–3410. doi: 10.1111/j.1572-0241.2001.05342.x

Siringo, S., Vaira, D., Menegatti, M., Piscaglia, F., Sofia, S., Gaetani, M., et al. (1997). High prevalence of *Helicobacter pylori* in liver cirrhosis: relationship with clinical and endoscopic features and the risk of peptic ulcer. *Dig. Dis. Sci.* 42, 2024–2030. doi: 10.1023/A:1018849930107

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

Stenn, F. (1960). Cave disease or speleonosis. *Arch. Intern. Med.* 105, 181–183. doi: 10.1001/archinte.1960.00270140003001

Sun, Y. Z., Zhang, D. H., Cai, S. B., Ming, Z., Li, J. Q., and Chen, X. (2018). MDAD: a special resource for microbe-drug associations. *Front. Cell. Infect. Microbiol.* 8:424. doi: 10.3389/fcimb.2018.00424

Sze, M. A., Dimitriu, P. A., Hayashi, S., Elliott, W. M., McDonough, J. E., Gosselink, J. V., et al. (2012). The lung tissue microbiome in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* 185, 1073–1080. doi: 10.1164/rccm.201111-2075OC

Tamboli, C. P., Neut, C., Desreumaux, P., and Colombel, J. F. (2004). Dysbiosis in inflammatory bowel disease. *Gut* 53, 1–4. doi: 10.1136/gut.53.1.1

Thiele, I., Heinken, A., and Fleming, R. M. (2013). A systems biology approach to studying the role of microbes in human health. *Curr. Opin. Biotechnol.* 24, 4–12. doi: 10.1016/j.copbio.2012.10.001

Torgerson, P. R., and Mastroiacovo, P. (2013). The global burden of congenital toxoplasmosis: a systematic review. *Bull. World Health Organ.* 91, 501–508. doi: 10.2471/BLT.12.111732

Trifan, A., Stoica, O., Stanciu, C., Cojocariu, C., Singeap, A. M., Girleanu, I., et al. (2015). *Clostridium difficile* infection in patients with liver disease: a review. *Eur. J. Clin. Microbiol. Infect. Dis.* 34, 2313–2324. doi: 10.1007/s10096-015-2501-z

Vael, C., Vanheirstraeten, L., Desager, K. N., and Goossens, H. (2011). Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol.* 11:68. doi: 10.1186/1471-2180-11-68

Van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

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.e941–943. doi: 10.1016/j.jaci.2011.07.027

Walker, A. W., Sanderson, J. D., Churcher, C., Parkes, G. C., Hudspith, B. N., Rayment, N., et al. (2011). High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. *BMC Microbiol.* 11:7. doi: 10.1186/1471-2180-11-7

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Wang, X., Heazlewood, S. P., Krause, D. O., and Florin, T. H. (2003). Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis. *J. Appl. Microbiol.* 95, 508–520. doi: 10.1046/j.1365-2672.2003.02005.x

Watson, A. J., and Collins, P. D. (2011). Colon cancer: a civilization disorder. *Dig. Dis.* 29, 222–228. doi: 10.1159/000323926

Weber, J., Illi, S., Nowak, D., Schierl, R., Holst, O., Von Mutius, E., et al. (2015). Asthma and the hygiene hypothesis. Does cleanliness matter? *Am. J. Respir. Crit. Care Med.* 191, 522–529. doi: 10.1164/rccm.201410-1899OC

Yao, Q., Wu, L., Li, J., Yang, L. G., Sun, Y., Li, Z., et al. (2017). Global Prioritizing disease candidate lncRNAs via a multi-level composite network. *Sci. Rep.* 7:39516. doi: 10.1038/srep39516

Zhang, W., Qu, Q., Zhang, Y., Wang, W., Zhang, W., Qu, Q., et al. (2017). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhao, Y., Chen, X., and Yin, J. (2018). A novel computational method for the identification of potential miRNA-disease association based on symmetric non-negative matrix factorization and kronecker regularized least square. *Front. Genet.* 9:324. doi: 10.3389/fgene.2018.00324

Zhou, X., Menche, J., Barabasi, A. L., and Sharma, A. (2014). Human symptoms-disease network. *Nat. Commun.* 5:4212. doi: 10.1038/ncomms5212

Zhu, Q., Jin, Z., Wu, W., Gao, R., Guo, B., Gao, Z., et al. (2014). Analysis of the intestinal lumen microbiota in an animal model of colorectal cancer. *PLoS ONE* 9:e90849. doi: 10.1371/journal.pone.0090849

Conflict of Interest Statement: 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.

Copyright © 2019 Qu, Zhao and Yin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 04 March 2019
doi: 10.3389/fmicb.2019.00395

Artificial Neural Networks for Prediction of Tuberculosis Disease

Muhammad Tahir Khan[1,2†], Aman Chandra Kaushik[2†], Linxiang Ji[3], Shaukat Iqbal Malik[1], Sajid Ali[4] and Dong-Qing Wei[2*]*

[1] Department of Bioinformatics and Biosciences, Capital University of Science and Technology, Islamabad, Pakistan, [2] College of Life Sciences and Biotechnology, The State Key Laboratory of Microbial Metabolism, Shanghai Jiao Tong University, Shanghai, China, [3] Department of Physics, Thompson Rivers University, Kamloops, BC, Canada, [4] Provincial Tuberculosis Reference Laboratory, Hayatabad Medical Complex, Peshawar, Pakistan

Background: The global burden of tuberculosis (TB) and antibiotic resistance is attracting the attention of researchers to develop some novel and rapid diagnostic tools. Although, the conventional methods like culture are considered as the gold standard, they are time consuming in diagnostic procedure, during which there are more chances in the transmission of disease. Further, the Xpert MTB/RIF assay offers a fast diagnostic facility within 2 h, but due to low sensitivity in some sample types may lead to more serious state of the disease. The role of computer technologies is now increasing in the diagnostic procedures. Here, in the current study we have applied the artificial neural network (ANN) that predicted the TB disease based on the TB suspect data.

Methods: We developed an approach for prediction of TB, based on an ANN. The data was collected from the TB suspects, guardians or care takers along with samples, referred by TB units and health centers. All the samples were processed and cultured. Data was trained on 12,636 records of TB patients, collected during the years 2016 and 2017 from the provincial TB reference laboratory, Khyber Pakhtunkhwa, Pakistan. The training and test set of the suspect data were kept as 70 and 30%, respectively, followed by validation and normalization. The ANN takes the TB suspect's information such as gender, age, HIV-status, previous TB history, sample type, and signs and symptoms for TB prediction.

Results: Based on TB patient data, ANN accurately predicted the *Mycobacterium tuberculosis* (MTB) positive or negative with an overall accuracy of >94%. Further, the accuracy of the test and validation were found to be >93%. This increased accuracy of ANN in the detection of TB suspected patients might be useful for early management of disease to adopt some control measures in further transmission and reduce the drug resistance burden.

Conclusion: ANNs algorithms may play an effective role in the early diagnosis of TB disease that might be applied as a supportive tool. Modern computer technologies should be trained in diagnostics for rapid disease management. Delays in TB diagnosis and initiation treatment may allow the emergence of new cases by transmission, causing high drug resistance in countries with a high TB burden.

Keywords: ANN, TB, data, diagnosis, drug resistance

Abbreviations: ANN, artificial neural network; ATH, Ayub Teaching Hospital; ATO, Agency TB officers; CSF, cerebrospinal fluid; DTO, district TB control officers; KPK, Khyber Pakhtunkhwa; MMTH, Mufti Mehmood Teaching Hospital; MTB, *Mycobacterium tuberculosis*; PMDT, Programmatic Management of Drug Resistant TB; PTRL, Provincial tuberculosis control program; TB, tuberculosis.

OPEN ACCESS

Edited by:
Xing Chen,
China University of Mining and Technology, China

Reviewed by:
Feng Zhu,
Zhejiang University, China
Willy Ssengooba,
Makerere University, Uganda

****Correspondence:***
Shaukat Iqbal Malik
drshaukat@cust.edu.pk
Dong-Qing Wei
dqwei@sjtu.edu.cn

† 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: *12 January 2019*
Accepted: *14 February 2019*
Published: *04 March 2019*

Citation:
Khan MT, Kaushik AC, Ji L, Malik SI, Ali S and Wei D-Q (2019) Artificial Neural Networks for Prediction of Tuberculosis Disease. Front. Microbiol. 10:395. doi: 10.3389/fmicb.2019.00395

INTRODUCTION

According to the (World Health Organization [WHO], 2018), 1.7 billion people (23%) of the world's population are estimated to have latent TB infection, indicating a risk of developing active TB during their lifetime (World Health Organization [WHO], 2018). Approximately 10.4 million incidences of TB occurred worldwide, including 5 million (56%) men, 3.5 million (34%) women and 1 million (10%) among children (WHO, 2017). Due to an increase in the world's population, the health care units are continuously struggling to improve the standard and reduce the transmission and cost. Methods commonly used to diagnose TB include, GeneXpert assay, sputum-smear microscopy and chest radiography (Dheda et al., 2017; Ejeta et al., 2018). However, diagnosis became more complicated when the infectious agent spread to other parts of the body – this is referred to as extra pulmonary TB. All these diagnostic methods possess some limitations. Culture method is considered the gold standard for detection of the causative agent of TB, *Mycobacterium tuberculosis* (MTB) but it is time consuming in diagnosis and the chances of contamination are high (Crowle et al., 1991; Osman et al., 2010; Asgharzadeh et al., 2015). Some common issues reported from other diagnostics methods include performance issues, sputum samples from children (pediatric cases), live MTB, highly skilled medical staff for high throughput tools and high cost (Dheda et al., 2017; Pandey et al., 2017). Delay in diagnosis may lead to drug resistance, multidrug resistance (MDR), where an isolate shows resistance to two first line drugs, rifampicin and isoniazid, and extensive drug resistance (XDR) which include MDR and also show resistance to fluoroquinolones and at least one of the injectable drug (Seung et al., 2015).

In health sciences, wet lab tests can be time consuming and the chances of contamination could further lead the disease to an irreversible state. Although the Xpert MTB/RIF assay offers a fast diagnostic facility within 2 h, but due to low sensitivity in some sample types and cost may lead to more serious state of disease (Pandey et al., 2017).

In the last few decades, the researchers have collected an extensive amount of biological data in genomics, proteomics, and in some other fields of biology during the gene and protein expression analysis. To extract some meaningful information and interpret the results, high throughput computational algorithms have been developed (Fojnica et al., 2016; Dande and Samant, 2018). In bioinformatics, data mining is a process of extracting useful information deep inside of large datasets (Sebban et al., 2002; Zheng et al., 2008; Li et al., 2013). These techniques also involve artificial intelligence, statistics, machine learning, and visualization (Li et al., 2013; Dande and Samant, 2018). Such techniques are applied to expose and analyze the hidden information inside the data or sometimes also called Intelligent Data Analysis (IDA), for better prediction of results. This knowledge discovery obtained from health data has some major objectives, including diagnosis in health sciences and simulations (Mello et al., 2006; Guillet and Hamilton, 2007; Chang et al., 2012).

Traditionally implemented diagnostic methods for tuberculosis patients can be minimized with data mining approaches. National and International laboratory researchers are currently involved in developing new diagnostics, and their evaluation helps to introduce more rapid and accurate methods in the diagnosis of TB along with the evaluations of alternative algorithms for TB reference laboratories (Parsons et al., 2011).

Artificial Neural Networks (ANNs) are operated by using algorithms to interpret non-linear data, independent of sequential pattern. The networks consist of a number of smaller units called neurons, organized between the input of data and the output of results into many layers. The ANN perform and behave like biological neurons, and this behavior may be learned through a backpropagation process. In this process, the precise output of a data set is previously known as input into the network. The least mean square difference of the entire data set is minimized by the continuous comparison of output of the ANN to the known output. A good level precision is adjusted by performing complex tasks without many computing resources (Drew and Monson, 2000).

The main advantage usually provided by ANN is their capability to extract hidden linear and non-linear relationships, even in the high dimensional and complex data sets (Zhang et al., 2010). In order to ease clinical decision makers, some more rapid evaluation techniques with low costs and good precision may further support in the diagnosis of TB to give optimum time for therapy, especially in TB high burden countries. Modern methods in data mining along with some traditional methods like regression have proven to be useful for comparison of prediction power of different models. The objective behind the current study is to support physicians in diagnosis, using predictive models as a diagnostic method for TB. Here, this investigation presents the data mining methods, i.e., classification, decision tree algorithm on the TB suspect data sets with selected attributes of patients to predict the presence or absence of TB disease. This information can be applied to develop less expensive diagnostic methods, dropping drug effects, data modeling, management of health care information systems, public health, and also patient's future prediction.

MATERIALS AND METHODS

Ethics Statement

This study proposal was approved and permitted by Institutional Committee (Ref 30/CUST2017) and incharge and molecular biologist, Provincial Tuberculosis Research Lab (PTRL) (Ref No. 1-06-17) where individual patient names and sensitive information were removed and neither of the these have been linked with an individual TB suspect. Further, the study was also conducted according to the WHO Standards and Operational Guidance for Ethics Review of World Health Organization [WHO] (2011). Annex-3(IV)-B (13).

Data Mining of TB Patient Data

Data in the current research was retrieved from TB control program at PTRL, Hayatabad Medical Complex Peshawar. All follow up and diagnostic patients have been included and the data of patients has been collected from guardians or care takers.

TABLE 1 | Characteristics of TB suspects/patients received from KPK TB units.

Characteristics	No.	Characteristics	No.	Characteristics	No.
Gender		**Hiv status**		**Sample type**	
Female	6445	Hiv +ve	534	Sputum	11089
Male	6043	Hiv −ve	5	Pleural fluid	343
*Other	148	Unknown	12097	*BAL	312
Total	12636	Total	12636	Gastric aspirate	224
				*CSF	172
History		**Age group**		Pus	124
Diagnosis	10416	0.1–14	1072	Ascetic fluid	117
Follow up	2220	15–24	3474	Tissue	104
Total	12636	25–34	2438	Urine	69
		35–44	1678	Pericardial fluid	43
Disease type		45–54	1690	Synovial fluid	16
Extra pulmonary	1323	55–64	1336	Lymph node	11
Pulmonary	11313	65–74	661	Total	12636
Total	12636	75–84	222	**Culture**	
		85–94	55	MTB	1809
		95–104	8	No growth	10827
		100–105	2	Total	12636
		Total	12636		

CSF, cerebrospinal fluid; BAL, bronchoalveolar lavage; Other, she male.

The data include location, age, gender, sample type, history, HIV status. The data set contain information's from 36 different TB units of Khyber Pakhtunkhwa (KPK). The characteristics of data is given **Table 1**, **2**.

Dataset Development

Suspects referred by TB units and health care centers during the years 2016 and 2017, were included. Samples were processed and cultured according the previous study (Kent and Kubica, 1985; Khan et al., 2018) and MTB negative and positive was confirmed after the culture result. Further confirmation was carried, using BD MGIT MTBc identification test (TBc ID, Ref: 245159, Becton, Dickinson), a rapid chromatographic immunoassay which detects the MTB complex antigen MPT64 secreted during culture (Arora et al., 2015). The dataset was validated by MATLAB software (Attaway, 2013). Total 12,636 inputs were used in order achieve a good output efficiency, where 70% were used as a training dataset and the remaining 30% were used as testing (Drew and Monson, 2000; Kulkarni et al., 2017). The validation observed for test dataset was about 93.71%.

TABLE 2 | Number of TB suspects received from different units of KPK province.

S.No.	Health center	Cases	S.No.	Health center	Cases
1	*ATO Khyber agency	71	20	DTO Sawabi	32
2	*CMH	296	21	DTO Shangla	4
3	*DTO Bajawar agency	11	22	DTO Swat	6
4	DTO Bannu	8	23	Hayatabad medical complex	1903
5	DTO Buner	8	24	Khyber teaching hospital	699
6	DTO Charsadda	54	25	Kuwait hospital	21
7	DTO Chitral	5	26	Leady reading hospital	2
8	DTO D.I. Khan	8	27	Mardan medical complex	8
9	DTO Dir Lower	7	28	*PMDT ATH	627
10	DTO Dir Upper	13	29	PMDT leady reading hospital	5333
11	DTO Hangu	39	30	PMDT *MMTH	1255
12	DTO Kohat	8	31	PMDT Swat	855
13	DTO Kohistan	1	32	Private	2
14	DTO Lakki	2	33	Peshawar reference lab	111
15	DTO Malakand	4	34	Rehman medical institute	66
16	DTO Mansehra	1	35	DTO Takht Nusrati	4
17	DTO Mardan	17	36	TB control Ganj	1
18	DTO Noshera	37			
19	DTO Peshawar	1117		Total	12636

ATH, Ayub Teaching Hospital; ATO, Agency TB officers; CMH, Combined Military Hospital; DTO, district TB control officers; MMTH, Mufti Mehmood Teaching Hospital; PMDT, Programmatic Management of Drug Resistant TB.

Artificial Neural Networks Approach

Artificial neural networks are nature inspired algorithms (Fojnica et al., 2016) that include input layer node, hidden layer node, and output node. Every node in a layer has one parallel node in the layer following it, thereby consequently building the stacking. Back-propagation learning algorithm is based on gradient descent search algorithms to fiddle with the correlation weight (Sollich and Krogh, 1996; Kaushik and Sahi, 2018). The output of every neuron was the aggregation of information of neurons of the prior stage multiplied by parallel weights with biased value. Input value was transformed into output with respect to activated functions shown in **Figure 1**.

Step 1 – Normalization of MTB Dataset.

TB patients dataset was normalized according to proposed study (Kaushik and Sahi, 2018).

$$V_{new} = (V_{old} - MinV) / (MaxV - MinV) * (D_{max} - D_{min}) + D_{min}$$

where V_{new} represent new assessment post-normalization, V_{old} is the assessment before normalization, MinV is the variable's minimum assessment, MaxV is the variable's maximum estimation. Dmax and Dmin are the maximum estimation succeeding normalization and the minimum assessment subsequent to normalization, respectively.

Step 2 – Input the data for training, the interrelated values of input and output execute for training using feed forward back propagation neural network algorithms.

Step 3 – Set Network constraint.

Step 4 – Calculate the neurons of output, every neuron output signal calculated using

$$net_j = \sum_{i=1\sim m} w_{ji}\, x_i + b_j$$

where netj and w_{ji} are output neurons and connection weight neurons, respectively, while xi and bj are the input signal, and bias neurons. The sigmoid function or logistic function, also called the sigmoidal curve (Seggern, 2016), was used for netj and every neuron of ten hidden layers.

Step 5 – Signal of output layers' calculation using,

$$net_k = TV_k + \delta_K^L$$

where TV_k is target value of output neurons and δ_K^L is the error of neuron.

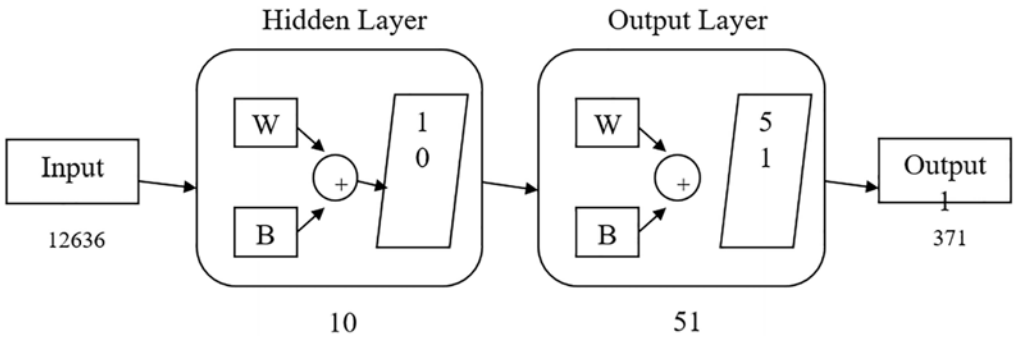


FIGURE 1 | Neural network structure of MTB dataset where dataset was categorized into three categories input, hidden, and output layers. "W" represent weight parameter with layer node, "B" represent bias unit.

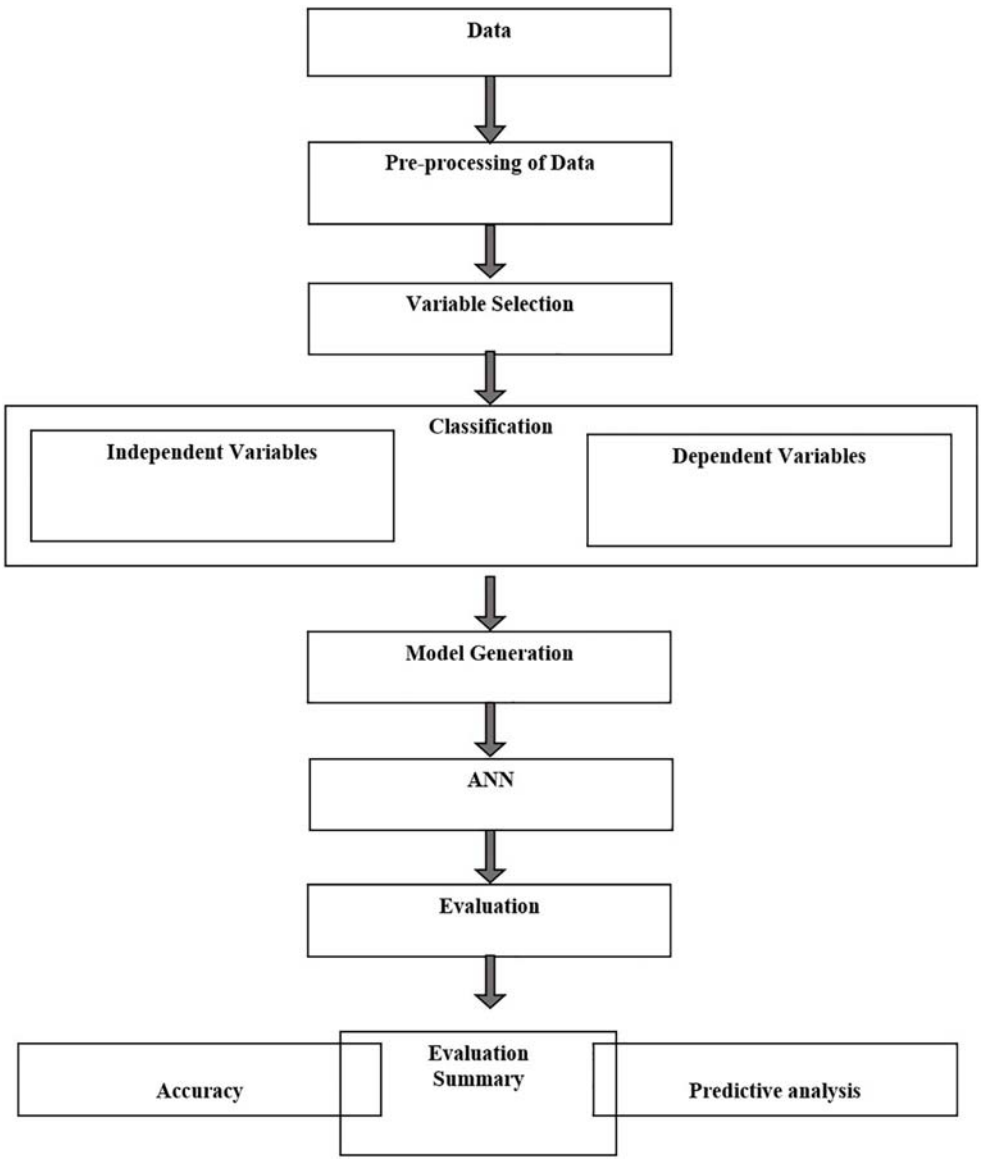


FIGURE 2 | Flow chart of ANN methodology for data processing, normalization, training, testing, and prediction.

Step 6 – Compute the error of neuron k and step 3 and step 6 were repetitive until network was congregated, and the error was computed using,

$$SSE = \sum_{i=1\sim n} (T_i - Y_i)^2$$

where T_i is actual assessment and Y_i is estimated assessment.

A step by step flowchart methodology has been given in **Figures 1**, **2**.

RESULTS

The drug resistance and patient's characteristics has been shown in **Figure 4**. MDR are very high among the population of KPK followed by other first line drug resistance. Although XDR are very few, they are very hard to treat and often take years to recover and the chances are very rare. Mono-resistant (resistant to any single drug) and poly resistant (resistant to any two or more drugs other than MDR and XDR) have been found in significant numbers. This high prevalence of drug resistance may be due to the delay type of diagnosis of some gold standard methods like culture. Owing to the current situation, we applied ANN on the patient records to find the accuracy of prediction.

Based on the ANN, the data used in the current study has 12,636 records (**Table 1**), where 70 and 30% were used as training and test sets, respectively. The accuracies of test and validation to predict TB based on patient data, were found at 93.90 and 93.71%, respectively. ANN-based, this algorithm accurately predicted that a suspect may have TB or not and generated the output through the hidden layer implementation. The hidden layers, learning parameters of ANN were as follows.

Number of input layer units	**12636**
Number of hidden layers	10
Number of first hidden layer units	10
Number of second hidden layer units	10
Number of output layer units	1
Momentum rate	0.88
Learning rate	0.70
Error after learning	0.000050
Learning cycle	30,000

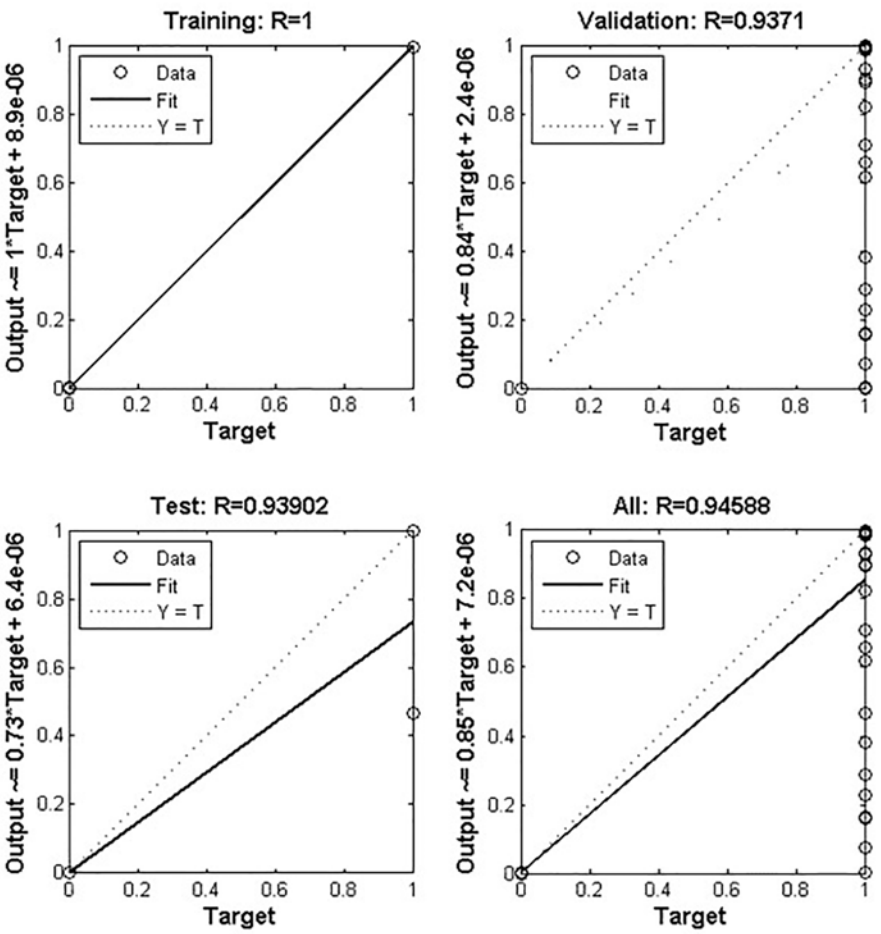


FIGURE 3 | Depicts artificial neural network prediction on the basis of normalized data of MTB. Out of 12636 records, 70% was training and 30% was test set where the validation score was achieved with an accuracy of about 94%. The overall model got an accuracy of 94.58%.

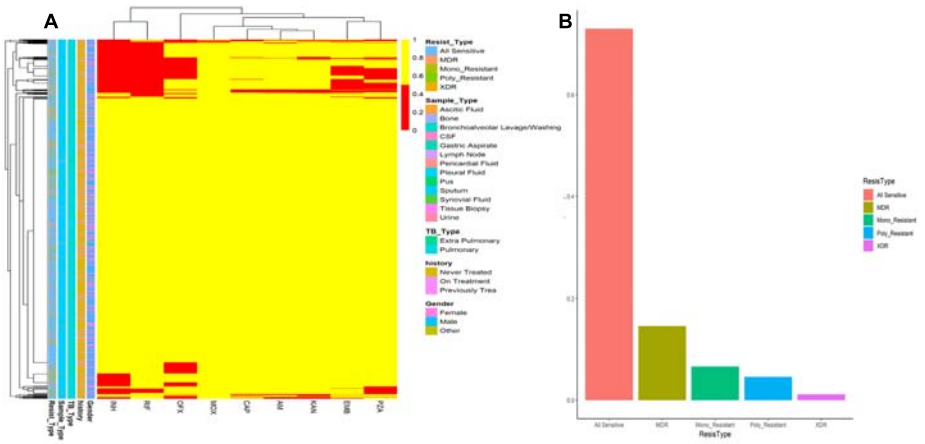


FIGURE 4 | (A) Drug resistance among MTB isolates from Khyber Pakhtunkhwa. Frequency of all sensitive, MDR, Mono_Resistant, Poly resistant, and XDR are color coded. **(B)** Patients data and drug resistance pattern. Drug resistance is shown by red color while drug sensitive isolates are colored yellow. Resistance type (Resist_Type), Sample_Type, TB_Type, previous TB History of patients, and Gender are color coded.

The Architecture of ANN

1. Initialize the weight and parameters μ ($\mu = 0.01$)
2. Compute the sum of the squared errors overall input $F(w) = e^T e$
   Where weight of network $w = [w1, w2, w3. \ldots w_n]$ and e is the error vector for the network.
3. Solve to obtain the increment of weight $\Delta w = [J^T J + \mu I]^{-1} J^T e$
   Where J is jacobian matrix, μ is learning rate neither μ is multiplied by decay rate $\beta(0 < \beta < 1)$.
4. Using $w + \Delta w$
   $F(w) < F(w)$ then (go back to Step 2)
   $W = w + \Delta w$
   $\mu = \mu.\beta$ ($\beta = 0.1$) (go back to Step 2)
   ELSE
   $\mu = \mu/\beta$ (go back to Step 2)
   END IF

The approach has been found efficient and possesses robust application for TB disease prediction with prediction accuracy. Using this approach, users can predict the active TB positively or negatively based on the patient's data after clinical sign symptoms including, cough that lasts 3 weeks or longer and pain in the chest coughing up blood or sputum (mucus from deep inside the lungs). Other symptoms of TB disease may include; weakness or fatigue, weight loss, no appetite, chills, fever, sweating at night (CDC Tuberculosis (TB), 2019).

User input data include; age, gender (male, female, other), sample type [bone, bone marrow, bronchoalveolar lavage, cerebrospinal fluid (CSF), gastric aspirate, lung biopsy, lymph nodes, pericardial fluid, pleural fluid, synovial fluid, pus, tissue biopsy, urine, sputum], history (follow up, diagnostic), HIV status (positive, negative, unknown). The model includes 70% training and 30% test set of the entire data set (12636 records) where the validation score was achieved with an accuracy of 94%.

The approach was written in MATLAB script where prediction accuracy was achieved as >94% based on ANN (**Figure 3**), dependent on dataset, where dataset input and hidden layer were categorized on two basic parameters W (weight) and B (bias unit), which contain ten sub-models and generate single output based on dataset accuracy. Users can predict their TB risk after entering their data, history, and the appearance of signs and symptoms.

DISCUSSION

Tuberculosis is a challenging disease; in spite of advanced technologies, the diagnosis is often difficult because of the nature of the disease (Dheda et al., 2017; WHO, 2017; World Health Organization [WHO], 2018). Clinical diagnosis requires standardization, where immunodiagnostic tests may help to improve sensitivity, but not in latent TB and some lack specificity (Newton et al., 2008; Elhassan et al., 2016). Xpert MTB/RIF have been saving our time to detect MTB, but decades old technologies like culture still remained the standard. Today, the battle against TB still poses one of the primary diagnostic problems in the pediatric laboratory (Dunn et al., 2016). Delay in notification and a weak coordination among TB management might be a cause to unnecessary diagnosis and treatment initiation (Yagui et al., 2006; Htun et al., 2018). Although the Xpert MTB/RIF assay offers fast diagnostic facility within 2 h, in some sample types like lymph node tissue biopsy (extrapulmonary TB) the overall sensitivity to rule out the TB is suboptimal (Creswell et al., 2014; Pandey et al., 2017; Tadesse et al., 2018). Performance was found to vary according to specimen type and acid-fast bacilli smear status. Further, the gold standard for MTB drug susceptibility testing is still culture on solid media, taking weeks to months to grow (Lin and Desmond, 2014; Dookie et al., 2018; Koch et al., 2018). Treatment is often empirical and initiated after looking at factors like past medical or social history, or the prevalence of drug resistance in that locality. These may delay the initiation of proper TB treatment that lead to drug resistance (Schaberg et al., 1996; Melchionda et al., 2013; Dookie et al., 2018; Khan et al., 2018). The high prevalence of drug resistance in TB high burden

countries may delay the initiation of appropriate treatment due to culture of MTB *in vitro* which is a time consuming method. These issues should be addressed through more studies especially in TB high burden regions. We found an increased drug resistance in the recent years' data (**Figure 4**) calculation, pointing toward advanced computational technologies to integrate for diagnosis in MTB prediction (Khan et al., 2018).

Modern neural networks have attained a great significance and importance in the recognition of images (Drew and Monson, 2000; Krizhevsky et al., 2012; Esteva et al., 2017; Kaushik and Sahi, 2018), speech recognition (Hinton et al., 2012), and processing of natural language (Socher et al., 2011). Medicinal researchers have started to apply these tactics in personalized clinical care. Diabetic retinopathy has also been identified through the approaches of deep neural networks (Gulshan et al., 2016) and classifying cancers of skin (Esteva et al., 2017). The applications of such approaches have also been found to be successful in computational biology and bioinformatics such as in inferring target gene expression (Chen et al., 2016), predicting RNA-binding protein sites (Zhang et al., 2016), and in identification and prediction of biomarkers for human chronological age (Putin et al., 2016). To reduce cost and time wastage, various data mining approaches may be helpful in diagnosis and on time initiation of TB therapy.

According to the WHO global TB report, 2018, India, Indonesia, China, Philippines, and Pakistan are the top five countries with 56% TB prevalence of the world. Timely TB diagnosis to reduce transmission and initiation of treatment to improve the outcomes for TB patients is essential, especially in high burden countries (Yagui et al., 2006; Dheda et al., 2017).

Classification and clustering algorithms are working efficiently with good precision in the prediction of the tuberculosis diagnosis. Presence of MTB and patient's data may support such model up to large extents. When handling high-dimensional classification problems, different modeling approaches may be used. Earlier works have applied multivariate logistic regression (Wisnivesky et al., 2005; Solari et al., 2008), classification trees (Mello et al., 2006; Aguiar et al., 2012) and ANN (Aguiar et al., 2013; Dande and Samant, 2018) for predicting smear-negative TB.

CONCLUSION

Artificial neural networks may be applied as a diagnostic tool for TB prediction and supportive in expanding the role for computer technologies in diagnostics for a rapid management of TB. Therefore, this high correlation (>94% accuracy) with the experimental result of MTB detection may help to choose optimal therapeutic regimens, especially in TB high burden countries. Delays in TB diagnosis and initiation of treatment may allow the emergence of new cases by transmission, and is one of the causes of high drug resistance in TB high burden countries.

The approach developed here may offer and support the rapid diagnosis of MTB with further additions such as drug resistance prediction in near future for better TB management.

DATA AVAILABILITY

The datasets in the present study will be provided upon reasonable request to the corresponding author.

AUTHOR CONTRIBUTIONS

Manuscript was designed by MK, DW, SM, and LJ. ANN was written and run by AK. Data analysis and manuscript writing were carried out by MK, SA, SM, and AK. Manuscript was revised by DW, LJ, and AK.

FUNDING

The work was supported by the grants from the Key Research Area grant 2016YFA0501703 of the Ministry of Science and Technology of China, the National Natural Science Foundation of China (Contract Nos. 61832019 and 61503244), the State Key Lab on Microbial Metabolism, and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2017ZD14). The simulations in this work were supported by the Center for High Performance Computing, Shanghai Jiao Tong University and also Higher Education Commission Islamabad, Pakistan under IRSIP No: 1-8/HEC/HRD/2017/8392.

ACKNOWLEDGMENTS

The present study was supported by SA and Anwar Sheed Khan, Molecular biologist microbiologist and project director PTRL Peshawar. Deborah L. Devis, University of Adelaide-Shanghai Jiao Tong University Joint Centre for Agriculture and Health, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Urrbrae, SA, Australia.

REFERENCES

Aguiar, F., Pinto, J., Torres, R., Seixas, J., and Mello, F. (2013). Artificial neural network (ANN) for prediction of pulmonary tuberculosis in hospitalized patients. *Eur. Respir. J.* 42:4469.

Aguiar, F. S., Almeida, L. L., Ruffino-Netto, A., Kritski, A. L., Mello, F. C., and Werneck, G. L. (2012). Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. *BMC Pulm. Med.* 12:40. doi: 10.1186/1471-2466-12-40

Arora, J., Kumar, G., Verma, A. K., Bhalla, M., Sarin, R., and Myneedu, V. P. (2015). Utility of MPT64 antigen detection for rapid confirmation of *mycobacterium tuberculosis* complex. *J. Glob. Infect. Dis.* 7, 66–69. doi: 10.4103/0974-777X.154443

Asgharzadeh, M., Kafil, H. S., Ghorghanlu, S., and Pourostadi, M. (2015). Laboratory cross-contamination of *Mycobacterium tuberculosis* in Northwest of Iran. *Egypt. J. Chest Dis. Tuberc.* 64, 665–669. doi: 10.1016/j.ejcdt.2015.03.018

Attaway, S. (2013). *Matlab: A Practical Introduction to Programming and Problem Solving*. Amsterdam: Elsevier, doi: 10.1016/B978-0-12-405876-7.00014-16

CDC Tuberculosis (TB) (2019). *Disease: Symptoms and Risk Factors |Features| CDC*. Available at: https://www.cdc.gov/features/tbsymptoms/index.html [Accessed September 22, 2018].

Chang, J., Arbeláez, P., Switz, N., Reber, C., Tapley, A., Davis, J. L., et al. (2012). Automated tuberculosis diagnosis using fluorescence images from a mobile microscope. *Med. Image Comput. Comput. Assist. Interv.* 15, 345–352.

Chen, Y., Li, Y., Narayan, R., Subramanian, A., and Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics* 32, 1832–1839. doi: 10.1093/bioinformatics/btw074

Creswell, J., Codlin, A. J., Andre, E., Micek, M. A., Bedru, A., Carter, E. J., et al. (2014). Results from early programmatic implementation of Xpert MTB/RIF testing in nine countries. *BMC Infect. Dis.* 14:2. doi: 10.1186/1471-2334-14-12

Crowle, A. J., Dahl, R., Ross, E., and May, M. H. (1991). Evidence that vesicles containing living, virulent *Mycobacterium tuberculosis* or *Mycobacterium avium* in cultured human macrophages are not acidic. *Infect. Immun.* 59, 1823–1831.

Dande, P., and Samant, P. (2018). Acquaintance to artificial neural networks and use of artificial intelligence as a diagnostic tool for tuberculosis: a review. *Tuberculosis* 108, 1–9. doi: 10.1016/j.tube.2017.09.006

Dheda, K., Gumbo, T., Maartens, G., Dooley, K. E., McNerney, R., Murray, M., et al. (2017). The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *Lancet Respir. Med.* 5, 291–360. doi: 10.1016/S2213-2600(17)30079-6

Dookie, N., Rambaran, S., Padayatchi, N., Mahomed, S., and Naidoo, K. (2018). Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. *J. Antimicrob. Chemother.* 73, 1138–1151. doi: 10.1093/jac/dkx506

Drew, P. J., and Monson, J. R. T. (2000). Artificial neural networks. *Surgery* 127, 3–11. doi: 10.1067/msy.2000.102173

Dunn, J. J., Starke, J. R., and Revell, P. A. (2016). Laboratory diagnosis of *Mycobacterium tuberculosis* infection and disease in children. *J. Clin. Microbiol.* 54, 1434–1441. doi: 10.1128/JCM.03043-3015

Ejeta, E., Beyene, G., Bonsa, Z., and Abebe, G. (2018). Xpert MTB/RIF assay for the diagnosis of *Mycobacterium tuberculosis* and Rifampicin resistance in high human immunodeficiency virus setting in gambella regional state, southwest ethiopia. *J. Clin. Tuberc. Mycobact. Dis.* 12, 14–20. doi: 10.1016/j.jctube.2018.06.002

Elhassan, M. M., Elmekki, M. A., Osman, A. L., and Hamid, M. E. (2016). Challenges in diagnosing tuberculosis in children: a comparative study from Sudan. *Int. J. Infect. Dis.* 43, 25–29. doi: 10.1016/j.ijid.2015.12.006

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi: 10.1038/nature21056

Fojnica, A., Osmanoviæ, A., and Badnjeviæ, A. (2016). "Dynamical model of tuberculosis-multiple strain prediction based on artificial neural network," in *Proceedings of the 2016 5th Mediterranean Conference on Embedded Computing (MECO)*, (Piscataway, NJ: IEEE), 290–293. doi: 10.1109/MECO.2016.7525763

Guillet, F., and Hamilton, H. J. (eds) (2007). *Quality Measures in Data Mining*. Berlin: Springer-Verlag. doi: 10.1007/978-3-540-44918-8

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410. doi: 10.1001/jama.2016.17216

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal. Process. Mag.* 29, 82–97. doi: 10.1109/MSP.2012.2205597

Htun, Y. M., Khaing, T. M. M., Yin, Y., Myint, Z., Aung, S. T., Hlaing, T. M., et al. (2018). Delay in diagnosis and treatment among adult multidrug resistant tuberculosis patients in yangon regional tuberculosis center, myanmar: a cross-sectional study. *BMC Health Serv. Res.* 18:878. doi: 10.1186/s12913-018-3715-4

Kaushik, A. C., and Sahi, S. (2018). HOGPred: artificial neural network-based model for orphan GPCRs. *Neural. Comput. Appl.* 29, 985–992. doi: 10.1007/s00521-016-2502-2506

Kent, P. T., and Kubica, G. P. (1985). *Public Health Mycobacteriology: A Guide for the Level III Laboratory*. Washington, DC: US Department of Health and Human Services.

Khan, M. T., Malik, S. I., Ali, S., Sheed Khan, A., Nadeem, T., Zeb, M. T., et al. (2018). Prevalence of pyrazinamide resistance in khyber pakhtunkhwa, Pakistan. *Microb. Drug Resist.* 24, 1417–1421. doi: 10.1089/mdr.2017.0234

Koch, A., Cox, H., and Mizrahi, V. (2018). Drug-resistant tuberculosis: challenges and opportunities for diagnosis and treatment. *Curr. Opin. Pharmacol.* 42, 7–15. doi: 10.1016/j.coph.2018.05.013

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* NIPS'12, (Red Hook, NY: Curran Associates Inc.), 1097–1105.

Kulkarni, P., Londhe, S., and Deo, M. (2017). Artificial neural networks for construction management: a review. *Soft Comput. Civil Eng.* 1, 70–88. doi: 10.22115/scce.2017.49580

Li, X., Ng, S.-K., and Wang, J. T. L. (2013). *Biological Data Mining and Its Applications in Healthcare*, 1st Edn. New Jersey, NJ: World Scientific Publishing.

Lin, S.-Y. G., and Desmond, E. P. (2014). Molecular diagnosis of tuberculosis and drug resistance. *Clin. Lab. Med.* 34, 297–314. doi: 10.1016/j.cll.2014.02.005

Melchionda, V., Wyatt, H., Capocci, S., Garcia Medina, R., Solamalai, A., Katiri, S., et al. (2013). Amikacin treatment for multidrug resistant tuberculosis: how much monitoring is required? *Eur. Respir. J.* 42, 1148–1150. doi: 10.1183/09031936.00184312

Mello, F. C. Q., Bastos, L. G. V., Soares, S. L. M., Rezende, V. M., Conde, M. B., Chaisson, R. E., et al. (2006). Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study. *BMC Public Health* 6:43. doi: 10.1186/1471-2458-6-43

Newton, S. M., Brent, A. J., Anderson, S., Whittaker, E., and Kampmann, B. (2008). Paediatric tuberculosis. *Lancet Infect. Dis.* 8, 498–510. doi: 10.1016/S1473-3099(08)70182-70188

Osman, M. K., Mashor, M. Y., and Jaafar, H. (2010). "Detection of *mycobacterium tuberculosis* in Ziehl-Neelsen stained tissue images using Zernike moments and hybrid multilayered perceptron network," in *Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics*, (Piscataway, NY: IEEE), 4049–4055. doi: 10.1109/ICSMC.2010.5642191

Pandey, S., Congdon, J., McInnes, B., Pop, A., and Coulter, C. (2017). Evaluation of the GeneXpert MTB/RIF assay on extrapulmonary and respiratory samples other than sputum: a low burden country experience. *Pathology* 49, 70–74. doi: 10.1016/j.pathol.2016.10.004

Parsons, L. M., Somoskövi, Á, Gutierrez, C., Lee, E., Paramasivan, C. N., Abimiku, A., et al. (2011). Laboratory diagnosis of tuberculosis in resource-poor countries: challenges and opportunities. *Clin. Microbiol. Rev.* 24, 314–350. doi: 10.1128/CMR.00059-10

Putin, E., Mamoshina, P., Aliper, A., Korzinkin, M., Moskalev, A., Kolosov, A., et al. (2016). Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging* 8, 1021–1033. doi: 10.18632/aging.100968

Schaberg, T., Rebhan, K., and Lode, H. (1996). Risk factors for side-effects of isoniazid, rifampin and pyrazinamide in patients hospitalized for pulmonary tuberculosis. *Eur. Respir. J.* 9, 2026–2030. doi: 10.1183/09031936.96.09102026

Sebban, M., Mokrousov, I., Rastogi, N., and Sola, C. (2002). A data-mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. *Bioinformatics* 18, 235–243. doi: 10.1093/bioinformatics/18.2.235

Seggern, D. H. (2016). *CRC Standard Curves and Surfaces with Mathematica*. Boca Raton, FL: CRC, doi: 10.1201/b16001

Seung, K. J., Keshavjee, S., and Rich, M. L. (2015). Multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis. *Cold Spring Harb. Perspect. Med.* 5:a017863. doi: 10.1101/cshperspect.a017863

Socher, R., Lin, C. C. Y., Ng, A. Y., and Manning, C. D. (2011). "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th International Conference on International Conference on Machine Learning* ICML'11, (Washington, DC: Omni press), 129–136.

Solari, A., Laurà, M., Salsano, E., Radice, D., Pareyson, D., and CMT-Triaal Study Group (2008). Reliability of clinical outcome measures in charcot-marie-tooth disease. *Neuromuscul. Disord.* 18, 19–26. doi: 10.1016/j.nmd.2007.09.006

Sollich, P., and Krogh, A. (1996). "Learning with ensembles: how over-fitting can be useful," in *Advances in Neural Information Processing Systems*, eds D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Cambridge, MA: MIT Press), 190–196.

Tadesse, M., Abebe, G., Bekele, A., Bezabih, M., Yilma, D., Apers, L., et al. (2018). Xpert MTB/RIF assay for the diagnosis of extrapulmonary tuberculosis:

a diagnostic evaluation study. *Clin. Microbiol. Infect.* doi: 10.1016/j.cmi.2018.12.018 [Epub ahead of print].

WHO (2017). *WHO | Global Tuberculosis Report 2017*. Available at: http://www.who.int/tb/publications/global_report/en/ [Accessed November 12, 2017].

Wisnivesky, J. P., Bonomi, M., Henschke, C., Iannuzzi, M., and McGinn, T. (2005). Radiation therapy for the treatment of unresected stage I-II non-small cell lung cancer. *Chest* 128, 1461–1467. doi: 10.1378/chest.128.3.1461

World Health Organization [WHO] (2018). *Global Tuberculosis Report 2018*. Geneva: World Health Organization.

World Health Organization [WHO] (2011). *Standards and operational guidance for ethics review of health-related research with human participants*. Geneva : World Health Organization. Available at: http://www.who.int/iris/handle/10665/44783

Yagui, M., Perales, M. T., Asencios, L., Vergara, L., Suarez, C., Yale, G., et al. (2006). Timely diagnosis of MDR-TB under program conditions: is rapid drug susceptibility testing sufficient? *Int. J. Tuberc. Lung Dis.* 10, 838–843.

Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., et al. (2016). A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.* 44, e32. doi: 10.1093/nar/gkv1025

Zhang, X., Pan, F., and Wang, W. (2010). "Finding high-order correlations in high-dimensional biological data," in *Link Mining: Models, Algorithms, and Applications*, eds P. S. Yu, J. Han, and C. Faloutsos (New York, NY: Springer), 505–534. doi: 10.1007/978-1-4419-6515-8_19

Zheng, H., Chruszcz, M., Lasota, P., Lebioda, L., and Minor, W. (2008). Data mining of metal ion environments present in protein structures. *J. Inorg. Biochem.* 102, 1765–1776. doi: 10.1016/j.jinorgbio.2008.05.006

Conflict of Interest Statement: 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.

Copyright © 2019 Khan, Kaushik, Ji, Malik, Ali and Wei. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

METHODS
published: 12 March 2019
doi: 10.3389/fmicb.2019.00428

DMSC: A Dynamic Multi-Seeds Method for Clustering 16S rRNA Sequences Into OTUs

Ze-Gang Wei[1,2] *and Shao-Wu Zhang*[1]*

[1] *Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an, China,* [2] *Institute of Physics and Optoelectronics Technology, Baoji University of Arts and Science, Baoji, China*

Next-generation sequencing (NGS)-based 16S rRNA sequencing by jointly using the PCR amplification and NGS technology is a cost-effective technique, which has been successfully used to study the phylogeny and taxonomy of samples from complex microbiomes or environments. Clustering 16S rRNA sequences into operational taxonomic units (OTUs) is often the first step for many downstream analyses. Heuristic clustering is one of the most widely employed approaches for generating OTUs. However, most heuristic OTUs clustering methods just select one single seed sequence to represent each cluster, resulting in their outcomes suffer from either overestimation of OTUs number or sensitivity to sequencing errors. In this paper, we present a novel dynamic multi-seeds clustering method (namely DMSC) to pick OTUs. DMSC first heuristically generates clusters according to the distance threshold. When the size of a cluster reaches the pre-defined minimum size, then DMSC selects the multi-core sequences (MCS) as the seeds that are defined as the n-core sequences ($n \geq 3$), in which the distance between any two sequences is less than the distance threshold. A new sequence is assigned to the corresponding cluster depending on the average distance to MCS and the distance standard deviation within the MCS. If a new sequence is added to the cluster, dynamically update the MCS until no sequence is merged into the cluster. The new method DMSC was tested on several simulated and real-life sequence datasets and also compared with the traditional heuristic methods such as CD-HIT, UCLUST, and DBH. Experimental results in terms of the inferred OTUs number, normalized mutual information (NMI) and Matthew correlation coefficient (MCC) metrics demonstrate that DMSC can produce higher quality clusters with low memory usage and reduce OTU overestimation. Additionally, DMSC is also robust to the sequencing errors. The DMSC software can be freely downloaded from https://github.com/NWPU-903PR/DMSC.

Keywords: multi-seeds, dynamic update, clustering, operational taxonomic units, 16S rRNA

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Lin Wan,
Academy of Mathematics and Systems Science (CAS), China
FangXiang Wu,
University of Saskatchewan, Canada

***Correspondence:**
Shao-Wu Zhang
zhangsw@nwpu.edu.cn

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

Received: *05 January 2019*
Accepted: *19 February 2019*
Published: *12 March 2019*

Citation:
Wei Z-G and Zhang S-W (2019) DMSC: A Dynamic Multi-Seeds Method for Clustering 16S rRNA Sequences Into OTUs. Front. Microbiol. 10:428. doi: 10.3389/fmicb.2019.00428

Abbreviations: AL, average linkage; MCC, matthews correlation coefficient; MCS, multi-core sequences; OTU, operational taxonomic units; rRNA, ribosomal RNA; std, standard deviations.

INTRODUCTION

Bacteria are the most diverse domain on our planet and play an essential role in various biogeochemical activities as well as an important role in human health and disease (Fuks et al., 2018). Characterizing the taxonomic community composition taken from an environmental sample is critical for understanding the bacterial world (Lane et al., 1985; Wei et al., 2016). Most of our knowledge about the microbial community descriptions comes from the 16S rRNA (ribosomal RNA) marker genes generated by high-throughput sequencing technology (Koslicki et al., 2013). Bypassing the necessity of isolating single organisms for cultivation, the advanced sequencing technology can produce millions of 16S rRNA and has become a powerful tool for in-depth analysis of bacterial community composition (Zhang et al., 2013; Wei and Zhang, 2018).

Usually, a fundamental first step for rapidly processing the 16S sequencing data is to cluster them into the OTUs (Turnbaugh et al., 2007; Peterson et al., 2009), which form the basis for estimating the species, diversity, composition, and richness of the microbes in the environment (Amir et al., 2017; Westcott and Schloss, 2017). Two major approaches for binning 16S rRNA sequences include: (i) taxonomy dependent methods, where each query sequence is compared against a reference taxonomy database and assigned to the organism of the best-matched annotated sequence using sequence searching (Altschul et al., 1990) or classification (Liu et al., 2017, 2018), and (ii) taxonomy independent methods (also called *de novo* clustering) (Chen et al., 2013b), where sequences are grouped into OTUs based on pairwise sequence similarities. However, a significant portion of microbes in a sample is contributed by unknown taxa which are not recorded in databases, thus taxonomy dependent methods are inherently limited by the completeness of reference databases (Chen et al., 2016). In contrast, *de novo* clustering methods divide sequences into OTUs without needing any reference database and have become the preferred choice for researchers (Cai et al., 2017).

In the past decade, a wide variety of *de novo* clustering methods has been proposed for binning OTUs. These methods can be further categorized into hierarchical clustering, heuristic clustering, model-based and network-based methods (Wei et al., 2017). Hierarchical clustering methods [e.g., mothur (Schloss et al., 2009), HPC-CLUST (Matias Rodrigues and von Mering, 2013), ESPRIT (Sun et al., 2009), and mcClust (Cole et al., 2013)] require a distance matrix derived either from all pairs sequences alignment or a multiple sequence alignment, then build a hierarchical tree with a predefined threshold to assign sequences into OTUs. Network-based methods [e.g., M-pick (Wang et al., 2013) and DMclust (Wei et al., 2017)] first construct a fully connected graph by computing all pairwise sequences distances and then employ the strategy of modularity community detection to generate OTUs. As a result, the computational complexity of both hierarchical and network-based methods is $O(N^2)$, where N is the number of sequences (Wei and Zhang, 2017; Wei et al., 2017). Model-based methods [e.g., CROP (Hao et al., 2011) and BEBaC (Cheng et al., 2012)] mainly apply some statistical model (e.g., Bayesian model) or mathematics framework (e.g., Gaussian mixture model) to describe sequence data then assign sequences to OTUs based on probability theory, and still, have a high computational burden (Chen et al., 2013a). Therefore, hierarchical clustering, model-based and network-based clustering methods quickly meet with the bottleneck in terms of computational time and memory usage for dealing with large-scale sequencing data (Wei et al., 2017).

A dozen of heuristic clustering methods such as CD-HIT (Li and Godzik, 2006), UCLUST (Edgar, 2010), DySC (Zheng et al., 2012), VSEARCH (Rognes et al., 2016), and DBH (Wei and Zhang, 2017) were developed to decrease the computational complexity. These methods build up clusters in an iterative incremental strategy. Each cluster is represented by one sequence (called seed) and each sequence is compared to all seeds. If the distance between one input sequence and a seed is within a given threshold, the input sequence is assigned to an existing cluster. Otherwise, this sequence becomes a seed of a new cluster. This procedure is repeated until all sequences are assigned. The computational complexity of heuristic clustering methods is $O(NM)$, where M is the number of seeds (usually $M \leq N$). Therefore, heuristic clustering methods run several orders of magnitude faster than other clustering algorithms and are more widely used in processing millions of 16S rRNA sequences (Cai and Sun, 2011).

Although heuristic clustering approaches are computationally efficient, they always overestimate the OTUs number and produce lower clustering quality than other methods (Huse et al., 2010; Wei and Zhang, 2015). Because most existing heuristic clustering methods just use one single sequence as the seed for each cluster, the results show an obvious sensitivity to the selected seeds that represent the clusters, especially when sequences datasets contain sequencing errors (Zheng et al., 2012; Chen et al., 2013a; Wei and Zhang, 2017). Therefore, selecting "good" seeds for one cluster is profoundly significant for heuristic clustering methods. In this work, inspired by the seed reselection procedure in DySC and the Gaussian model representation of one cluster in CROP, we proposed a **d**ynamic **m**ulti-**s**eeds **c**lustering (namely DMSC) method to pick OTUs. The DMSC algorithm consists of four main phases. First, heuristically generate clusters according to the distance threshold, which is similar to classical heuristic methods (e.g., CD-HIT or UCLUST). Second, when the size of a cluster reaches the pre-defined minimum size, select the MCS as seeds of a cluster, in which the distance between any two sequences is less than the distance threshold. Third, a new sequence is assigned to the corresponding cluster depending on the average distance to MCS and the distance standard deviation between each pairwise sequences in MCS. Finally, DMSC dynamically updates the MCS until no sequence is merged into the cluster.

Compared with other heuristic clustering methods, the unique characteristics of our DMSC method mainly manifest in the following three points. (i) DMSC selects MCS as the seeds in one cluster instead of the single seed representation used in most heuristic clustering methods such as CD-HIT and UCLUST; (ii) in DMSC, the MCS of one cluster is always dynamically updated with the cluster size increases, while the seed of each cluster in most other heuristic methods is always fixed; and (iii) according

to the average distance to MCS and the distance standard deviation between each pairwise sequences in MCS, a new sequence is assigned to the corresponding cluster, while other heuristic methods assign the new sequence to one cluster just base on the distance with the seed sequence. Four experimental results demonstrate that DMSC can achieve higher quality clusters and reduce OTU overestimation with low memory usage. Additionally, DMSC is also robust to sequencing errors.

MATERIALS AND METHODS

The first motivation of our DMSC method is to decrease the sensitivity of single seed representation to sequencing errors in most heuristic clustering methods. Here we select the MCS as seeds of a cluster, in which the distance between any two sequences is less than the distance threshold. There are two different parameters in DMSC approach: η (default value 25), the minimum sequence number in a cluster to ensure that the cluster contains enough sequences to yield a reliable MCS; and μ (default value 3), the time (multiple) of distance standard deviation between each pair of sequences in the MCS. These parameter settings have been evaluated in following experiments and the default values have robust performance. **Figure 1** is a flowchart showing the OTUs generating process with DMSC. It can be seen that DMSC method has four main phases: (i) according to the distance threshold θ, a series of clusters are formed by heuristic clustering of each sequence one by one; (ii) when the size of a cluster reaches the pre-defined minimum sequence number (η), the MCS is selected as the seeds; (iii) according to the average distance to MCS and the distance standard deviation (σ) between each pairwise sequences in MCS, a new sequence is assigned to the corresponding cluster; and (iv) after a new sequence is added to one cluster, update the MCS.

Generating Clusters

At the beginning of DMSC, the input sequences are sorted by abundance in a descending order. These can eliminate the influence of sequence input order on the clustering results. Then the first sequence is assigned to the first cluster and becomes the seed of this cluster. The second sequence is added to the cluster if the distance between the sequence and the seed is within the pre-defined threshold (θ), otherwise, this sequence is stored as a new seed for creating a new cluster. Repeat this process until the size of a cluster reaches the predefined threshold (η), then the MCS selection procedure is activated.

Selecting Multi-Core Sequences (MCS)

The multi-core sequences of one cluster is defined as the n-core sequences ($n \geq 3$), in which the distance between any two sequences within the cluster is less than the distance threshold (θ). If more than 3-core sequences are selected in the cluster, these core sequences are taken as seeds to represent this cluster, otherwise, one seed sequence is selected to represent this cluster. Although the MCS selection procedure can reduce OTU overestimation and decrease the sensitivity to the sequencing errors, it will increase the computational burden. Considering both the clustering quality and the computational burden, we select more than 3 core sequences (i.e., $n \geq 3$) as the seeds in this paper. The pseudo-code for the MCS selection procedure is outlined in the following **Figure 2**.

Assigning Sequences

One reason that heuristic clustering methods generally overestimate the OTUs number is that these methods just compare the distance with single seed to assign sequences. Model-based clustering methods can reduce OTU overestimation because they consider the distance distribution in one cluster. Therefore, we introduce the distance standard deviation (σ) between each pairwise sequences in one MCS in this work. That is:

$$\left|d(s, M_i)\right| \leq \mu^* \sigma_i \tag{1}$$

where M_i is the MCS of the i-th cluster, $d(s, M_i)$ is the average distance between sequence s and M_i, μ is the multiple constant, σ_i is the distance standard deviation of M_i. If the sequence s meets Equation 1, then s is merged into the i-th cluster. $d(s, M_i)$ and σ_i are defined as:

$$d\left(s, M_i\right) = \frac{1}{|M_i|} \sum_{i=1}^{|M_i|} d\left(s, s_i\right), \ s_i \in M_i \tag{2}$$

$$\sigma_i = \sqrt{\frac{1}{|M_i| - 1} \sum_{s_i, s_j \in M_i}^{s_i \neq s_j} \left[d\left(s_i, s_j\right) - \bar{d}_{M_i}\right]^2} \tag{3}$$

where $|M_i|$ is the sequence number in M_i, $\bar{d}_{M_i}$ is the average distance of all pairwise sequences in M_i.

Updating MCS

Once one sequence is merged into a cluster, the MCS will be updated according to the MCS selection procedure in **Figure 2**. Therefore, the MCS of one cluster is always dynamically updating with the cluster size increases.

After all the MCSs are no long change, all the isolated sequences are checked and assigned to the nearest neighbor clusters to form OTUs.

RESULTS

We compared our DMSC method with seven state-of-the-art OTUs clustering algorithms: CD-HIT (v.4.6.8) (Li and Godzik, 2006), UCLUST (v.11.0.667) (Edgar, 2010), DBH (Wei and Zhang, 2017), DySC (Zheng et al., 2012), ESPRIT-Forest (Cai et al., 2017), AL clustering algorithm implemented in mothur (v.1.40.5) (Schloss et al., 2009), and CROP (Hao et al., 2011). Among these methods, CD-HIT, UCLUST, DySC, and DBH are typical heuristic clustering approaches; mothur is an open source software package for analyzing the biological sequence data, and the AL clustering in mothur (mothur-AL) has been demonstrated that it is a reliable method to represent the actual distances between sequences (Westcott and Schloss, 2015); ESPRIT-Forest is a new parallel hierarchical clustering method, and CROP is

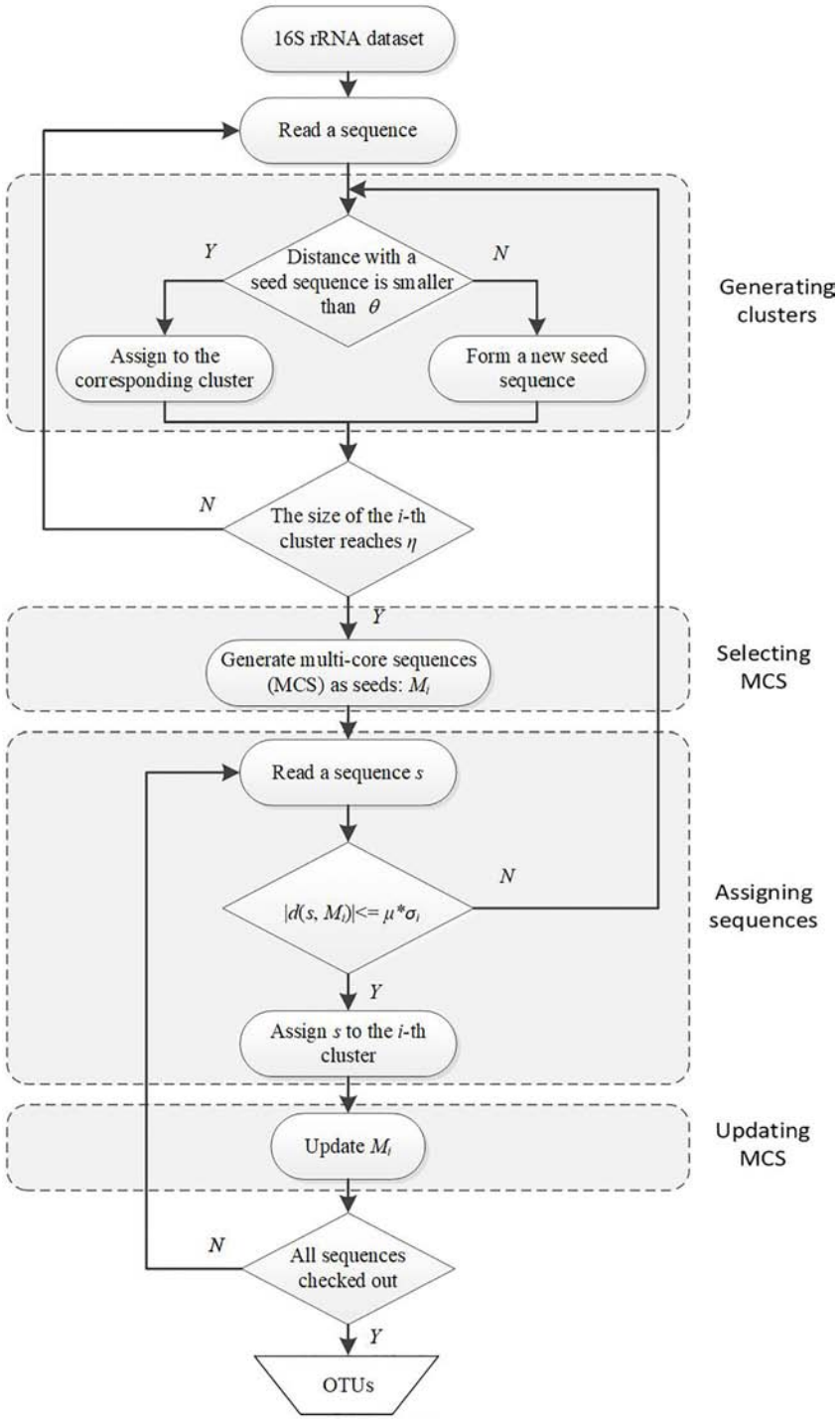


FIGURE 1 | Flowchart of DMSC algorithm. DMSC contains four main modules: generating clusters, selecting MCS, assigning sequences, and updating MCS. θ denotes the distance threshold; η is the minimum sequence number of a cluster to select the MCS; M_i is the MCS of the i-th cluster; σ_i is the distance standard deviation between each pairwise sequences in M_i; and μ denotes the multiple.

a model-based method. We conducted these methods on four benchmark datasets including two simulated dataset and three published real-life datasets. Some features of each benchmark dataset are shown in **Table 1**.

The metrics of OTUs number, NMI, and MCC are adopted to access the performance of every OTU picking method in the following experiments. The metrics of OTUs number and NMI have been widely used to compare the performance of OTU

Multi-core sequences selection
Input: cluster: *cluster* and distance threshold: θ
Output: multi-core sequences: *M*.
Multi-core sequences selection (*cluster*, θ)
1. Initialize: $k = 0$
2. For s_i in *cluster*:
3. $k = k + 1$
4. $mcs_k = \{ s_i \}$
5. For s_j in cluster ($s_j \neq s_i$):
6. If $d(s_j, s_m) \leq \theta, \ \forall s_m \in mcs_k$:
7. $mcs_k = mcs_k \cup s_j$
8. End
9. End
10. Return the mcs_i that has the maximum size ($\geqslant 3$).

FIGURE 2 | The pseudo-code of the MCS selection procedure for one cluster.

picking methods based on the known ground truth information datasets (Sun et al., 2011; Schmidt et al., 2015). Although the ground truth information (i.e., how many species the dataset includes, and what species the sequence belongs to) is always unknown for most real-life 16S rRNA sequencing dataset, it can be partially resolved by applying some searching methods against the reference database to annotate the 16S rRNA sequences (Cai and Sun, 2011; Chen et al., 2013b; Edgar, 2018). MCC metric was also used to evaluate the performance of OTU picking methods based on the sequence distance and clustering threshold without relying on an external reference (Schloss and Westcott, 2011), which is an objective metric to assess the clustering quality of OTUs picking methods (He et al., 2015; Westcott and Schloss, 2015; Schloss, 2016). The computational formulas of NMI and MCC are listed in **Supplementary File**.

All methods were executed on an Ubuntu 16.04.5 server with 16 3.2-GHz Intel Xeon (E5-2667V4) processors and 128 GB of RAM. And the running command lines of each method are listed in **Supplementary Table S1**.

Experiment 1: Stacked_60 Dataset

The Stacked_60 benchmark dataset was constructed by Barriuso et al. (2011), which is retrieved from 59 different bacterial genera in the NCBI and trimmed to obtain the V6 region (from positions 963 to 1063 in *E. coli*). Stacked_60 contains random mutation and is specially designed to test the accuracy of OTUs picking methods at different sequence distances. The taxa distance range and the taxa abundance are in 0.01–0.38 and 0.001–0.003, respectively.

Table 2 lists the maximum NMI value and the corresponding OTUs number, from which we can see that DMSC and CROP have higher maximum NMI value than the other methods, and different methods achieve the maximum NMI values at different distance thresholds. At the respective maximum NMI value, DMSC and CROP inferred 59 OTUs which equals to the expected number, while DBH, DySC, CD-HIT, mothur-AL and ESPRIT-Forest overestimated OTUs number, and UCLUST underestimated OTUs number.

Figure 3 shows the NMI values of DMSC, CROP, UCLUST, CD-HIT, DySC, DBH, mothur-AL, and ESPRIT-Forest with different distance thresholds on the Stacked_60 dataset. It can be seen that the NMI value of DMSC is almost identical to the CROP from 0.03 to 0.05 distance threshold, and also higher than that of other methods. In the range of 0.06∼0.09, DMSC achieved the highest NMI values, while the NMI value of CROP continuously drops, indicating that CROP is more sensitive to the distance threshold. Because the NMI values vary a lot in the range of 0.01∼0.02 distance thresholds for all methods, **Figure 3** just represents the NMI values from 0.03 to 0.10 distance thresholds. **Figure 4** depicts the MCC curves of eight methods with different distance thresholds on Stacked_60 dataset. From **Figure 4** we can see that DMSC method always achieved the highest MCC value in the range of 0.01∼0.10 distance thresholds. The NMI values, OTUs number and MCC values of eight methods in the range of 0.01∼0.1 distance thresholds can be found in **Supplementary Table S2**.

These results in **Figures 3**, **4**, **Table 1**, and **Supplementary Table S2** show that our DMSC method can accurately estimate the species number and obtain better cluster quality for Stacked_60 dataset.

Experiment 2: Simulated Dataset

We then considered another widely used simulated dataset to estimate the clustering accuracy, where the ground truths were directly taken from a simulator software (Cheng et al., 2012). A total of 22,000 sequences (∼500 bp) from 11 taxa were generated and each taxon contains 2,000 sequences with different substitution rates. Among these 11 taxa, three taxa are within 1% different from each other. Therefore, the expected OTUs number is 9.

TABLE 1 | Details of the benchmark datasets.

Datasets	Taxon number	Sequence number	Average length	Variable regions	Data source
Stacked_60 dataset	59	2,614	98 bp	V6	Barriuso et al., 2011
Simulated dataset	11	22,000	500 bp	–	Cheng et al., 2012
V6 dataset	177	∼310 K	121 bp	V6	Chen et al., 2013a
V4 dataset	68	∼511 K	253 bp	V4	Westcott and Schloss, 2015
Error datasets	30	150 K	120 bp	V6	Wei and Zhang, 2017

TABLE 2 | The maximum NMI values and OTUs number with different methods on stacked_60 dataset.

Methods	DMSC (0.03)	CROP (0.03)	DySC (0.03)	DBH (0.02)	CD-HIT (0.04)	UCLUST (0.09)	mothur-AL (0.06)	ESPRIT-Forest (0.08)
Max. NMI	0.99951	0.99951	0.99475	0.99868	0.99557	0.98528	0.96650	0.96614
OTUs	59	59	60	62	65	56	161	86

The value in the bracket is the distance threshold where each method achieves its maximum NMI. For mothur-AL method, the maximum NMI of mothur-AL is selected from the distance range of 0.01∼0.06 for reason that mothur-AL method just obtains the clustering results in these distance thresholds.

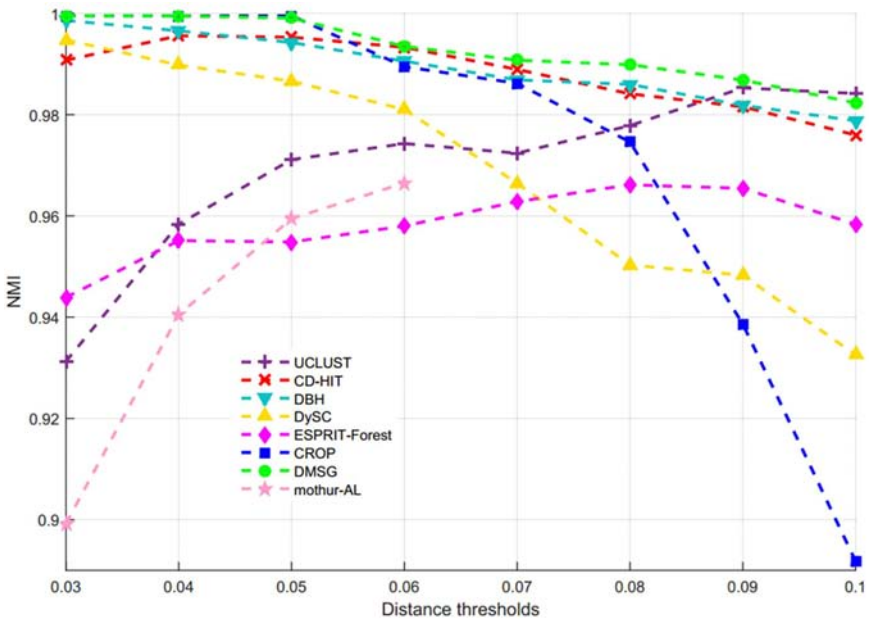


FIGURE 3 | NMI values of different algorithms on stacked_60 dataset.

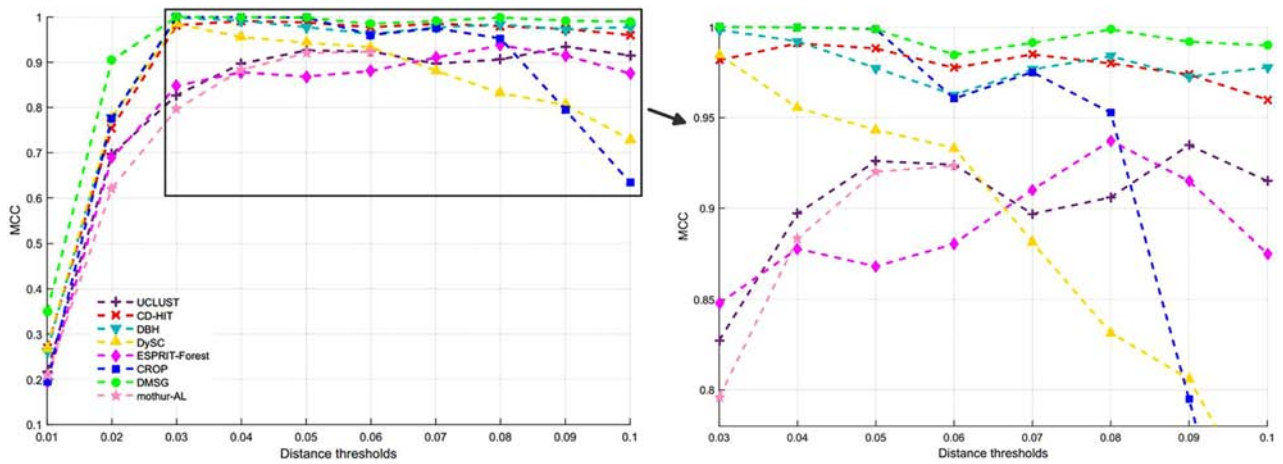


FIGURE 4 | MCC values of different methods on stacked_60 dataset.

By setting different distance thresholds ranging from 0.01 to 0.1, the maximum NMI values of seven methods at different distance thresholds and the corresponding inferred OTUs number are reported in **Table 3**, from which we can see that DMSC achieved the highest NMI (0.9503). Meanwhile, DMSC, CROP, DBH, and CD-HIT successfully obtained 9 OTUs at their best NMI value, while DySC, UCLUST, and ESPRIT-Forest overestimated OTUs. The NMI curves of seven methods are shown in **Figure 5**, from which we can see that DMSC achieved better NMI values than other methods at distance intervals

TABLE 3 | Maximum NMI values of seven methods on the simulated dataset.

Methods	DMSC (0.02)	CROP (0.03)	DySC (0.03)	DBH (0.03)	CD-HIT (0.05)	UCLUST (0.05)	ESPRIT-Forest (0.05)
Maximum NMI	0.9503	0.9334	0.9252	0.9293	0.9334	0.9107	0.8979
OTUs number	9	9	17	9	9	10	13

The value in the bracket is the distance threshold where each method achieves its maximum NMI.

[0.01, 0.04] and [0.07, 0.1], reaching the highest NMI value at 0.02 distance threshold; other methods obtained their best NMI values at different distance thresholds. **Figure 6** represents the MCC curve of seven methods with different distance thresholds ranging from 0.01 to 0.1, from which we can see that MCC values of DMSC are higher than that of other six methods in the range of 0.02∼0.07 distance thresholds. The NMI values, OTUs number and MCC values of seven methods are listed in **Supplementary Table S3**. These results indicate that DMSC has a better cluster performance than ESPRIT-Tree, CD-HIT, UCLUST, DBH, CROP, and DySC.

Experiment 3: V6 Variable Region Dataset From Human Gut Flora

In this experiment, we use one real-world benchmark dataset of the V6 variable region from human gut flora to evaluate the performance of OTUs picking methods. This dataset contains ∼310K sequences (average length: ∼121 bp) which are classified into 177 species and covers the V6 hypervariable region of 16S rRNA gene (Chen et al., 2013a). In order to reduce computational burden and remove statistical variations, each method was run 10 times and ∼30K reads were randomly extracted from the V6 dataset in each run.

Figure 7 describes the average NMI value as a function of the distance threshold over 10 runs for six methods, from which we can observe that DMSC has the highest NMI values than other methods in the range of 0.01∼0.07 distance thresholds, and DBH also achieved higher NMI values than CD-HIT, UCLUST, DySC, and ESPRIT-Forest from distance threshold interval [0.02, 0.08]. CD-HIT has the lowest NMI values except at 0.1 distance threshold. The average OTUs number inferred with six methods at different distance thresholds are described in **Supplementary Figure S1**, from which we can see that DMSC inferred fewer OTUs than CD-HIT, UCLUST, DBH and ESPRIT-Forest, but more than DySC at different distance thresholds. These can be explained by the fact that the sequence distance calculation in DySC is based on pairwise *k*-mer distances (Zheng et al., 2012), while other methods (including DMSC) are based on pairwise sequence alignment (PSA). It's reported that *k*-mer distance is looser than PSA (Sun et al., 2009). In other words, when setting to the same threshold (e.g., 0.03), more sequences of using the *k*-mer distance will satisfy the threshold to be clustered into one group, resulting in that DySC trends to generate fewer OTUs. However, DySC always gives less clustering accuracy and quality than DMSC in terms of the NMI (**Figure 7**) and MCC (**Figure 8**) evaluation metrics. **Supplementary Figure S2** reports the NMI std of six methods at different distance thresholds with 10 re-sampled runs, from which we can see that the NMI std of DMSC varies in the scope of 0.003∼0.012 at different distance thresholds. DMSC has the lowest std than other five methods in the range of 0.06∼0.09 distance thresholds and

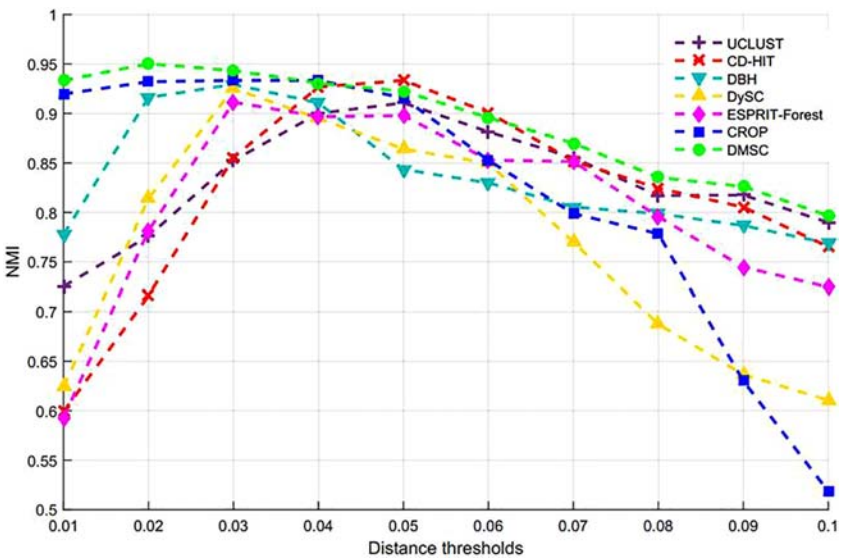


FIGURE 5 | NMI values of different methods on the simulated dataset.

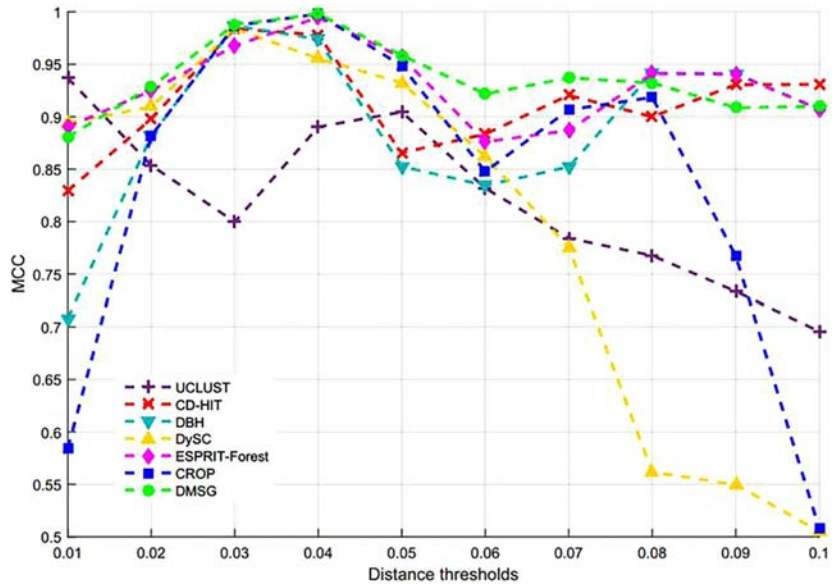


FIGURE 6 | MCC values of different methods on the simulated dataset.

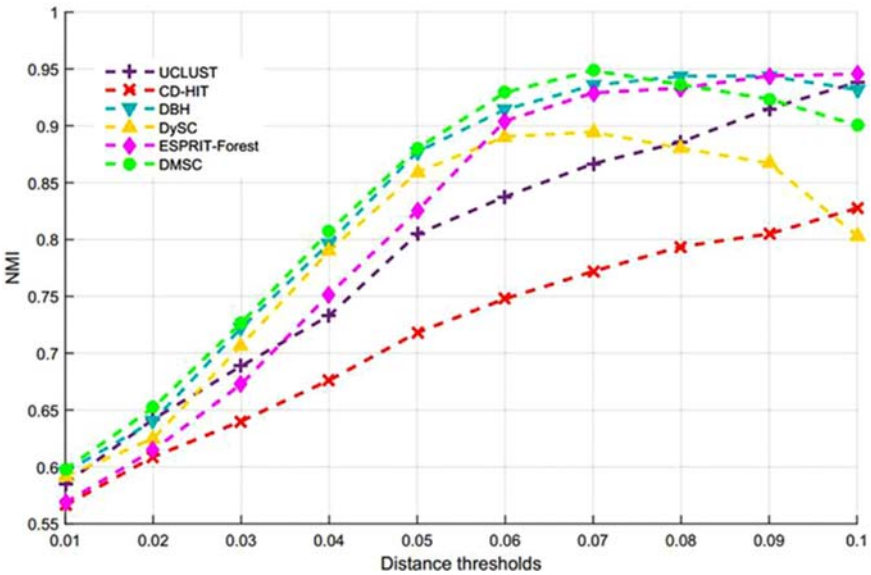


FIGURE 7 | Average NMI values of six methods at different distance thresholds on V6 data set.

almost equals to CD-HIT and UCLUST in the range of 0.01∼0.05 distance thresholds. **Figure 8** presents the MCC curves of six methods with different distance thresholds, from which we can see that the MCC values of DMSC and DBH are higher than that of other four methods in the range of 0.03∼0.10 distance thresholds. For reason that CROP takes longer running time to output the OTUs for the large-scale dataset, we did not list the results of CROP in this experiment. **Supplementary Table S4** lists the NMI values, OTUs number and MCC values of six methods, and **Supplementary Table S5** gives the *t*-test results of DMSC compared with the other four methods. These results in **Figures 7**, **8**, **Supplementary Figures S1**, **S2**, and **Supplementary Tables S4**, **S5** show that DMSC can generate the most robust estimations.

Experiment 4: V4 Variable Region Dataset From the Murine Gut

In this experiment, we adopt another real-world benchmark dataset of the V4 variable region from the Murine gut to assess the performance of OTUs picking methods. The V4 dataset was generated by Illumina's MiSeq platform

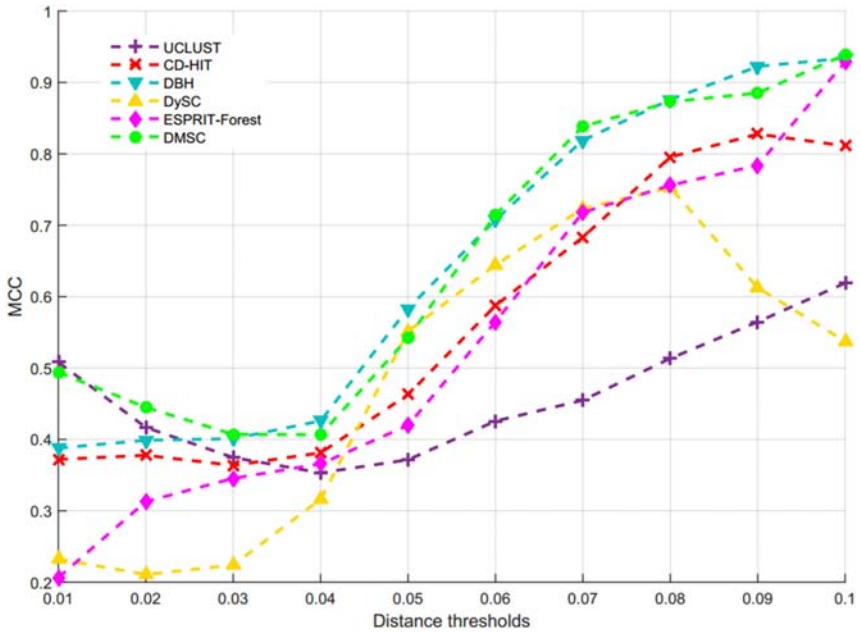


FIGURE 8 | The MCC values of six methods on V6 dataset.

(Westcott and Schloss, 2015), covering the V4 hypervariable region of 16S rRNAs from Murine microbiota [36]. The raw sequences of V4 dataset can be freely obtained from http:/www.mothur.org/MiSeqDevelopmentData/StabilityNoMetaG.tar. The ground-truth of V4 dataset can be extracted as followings. First, the pair end raw sequences were merged by FLASH (Magoč and Salzberg, 2011), then the usearch (Edgar, 2010) program was adopted to filter the merged sequences. Finally, the Python script (assign_taxonomy.py) in QIIME (Caporaso et al., 2010) was used to align the sequences for obtaining the ground-truth information with a stringent criterion. If the identity percentage is more than 97% (≥97%) and the length of the aligned region is more than 90% (≥90%) of the total length, the annotated sequences are retained. Thus, we obtained about ∼511K annotated reads, which were classified into 68 genera.

By setting different distance thresholds ranging from 0.01 to 0.15, the NMI curves of five methods are shown in **Figure 9**, and the inferred OTUs number of five methods at different distance threshold are presented in **Supplementary Figure S3**. **Figure 10** is the MCC curves of five methods at different distance thresholds. The NMI values, OTUs number and MCC values inferred with five methods at different distance thresholds are listed in **Supplementary Table S6**. Because DySC software returns a debug information, ESPRIT-Forest appears a segmentation fault (core dumped) information, and CROP is time-consuming on this large V4 dataset, we did not give the results of DySC, ESPRIT-Forest, and CROP in this experiment.

From **Figure 9**, we can see that most of NMI values of DMSC are higher than that of other four methods in the range of 0.01∼0.13 distance thresholds, and it is obviously higher than other three methods in the distance range of 0.09∼0.12. The results in **Supplementary Figure S3** show that DMSC and DBH inferred less OTUs than other methods, and DMSC inferred 67 OTUs which is near the ground truth at 0.09 distance threshold. From **Figure 10**, we can see that the MCC values of DMSC are higher than that of the other four methods except at 0.10 distance threshold. These results suggest that DMSC can achieve higher clustering quality than UCLUST, CD-HIT, DBH, and mothur-AL methods.

DISCUSSION

Inspired by the seed reselection strategy and model-based methods, we herein developed a novel dynamic multi-seeds heuristic method for picking OTUs from 16S rRNA sequences. Besides the distance threshold θ given by users, DMSC also needs another two parameters in picking OTUs procedure: η and μ. How these two parameters affect the clustering results needs to be further investigated. In the following, we tested the parameter effect on the simulated dataset used in experiment 2. We first tested the effect of the η by fixing μ (e.g., $\mu = 3$). The NMI values at different distance thresholds are presented in **Supplementary Figure S4**, from which we can see that we can see that the NMI values of $\eta = 10, 15, 20, 15$ in the range of 0.02∼0.1 distance thresholds are nearly equal, indicating that η has little influence on the clustering results. **Supplementary Figure S5** shows the effect of μ by fixing η (e.g., $\eta = 25$). From **Supplementary Figure S5**, we found that the NMI values of $\mu = 3, 4$ are higher than that of $\mu = 1, 2$ in the range of 0.01∼0.1 distance thresholds. Therefore, we select $\eta = 25$ and $\mu = 3$ as the default parameter values in our DMSC method.

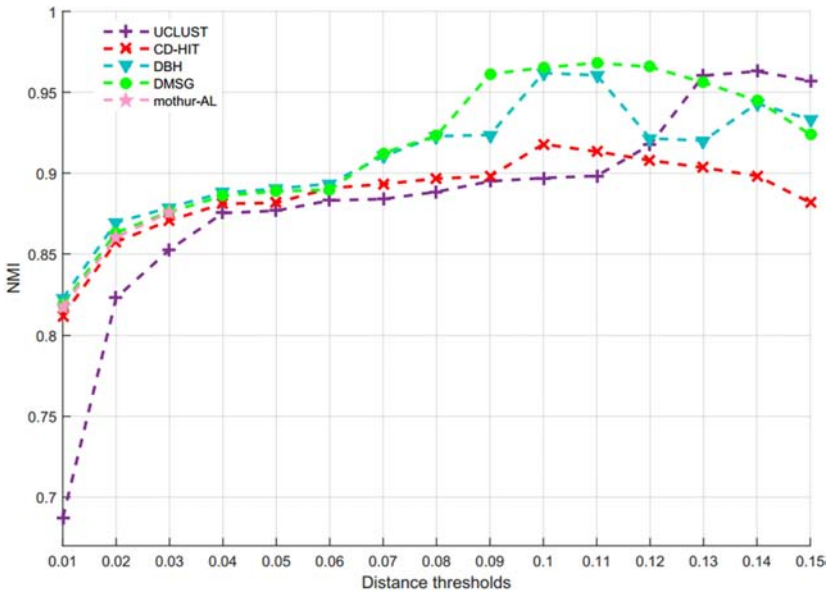


FIGURE 9 | NMI values of five methods at different distance thresholds on V4 dataset.

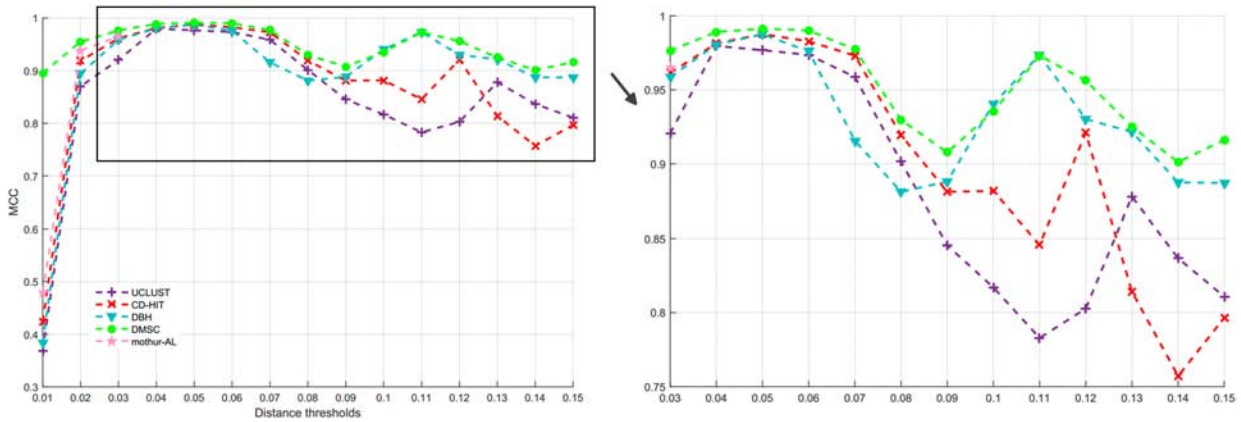


FIGURE 10 | MCC values of five methods at different distance thresholds on V4 dataset.

Sequencing errors (i.e., deletion, insertion, and substitution) are inevitably introduced during the high-throughput sequencing procedure, which can easily lead to OTUs overestimation (Schmidt et al., 2015). In order to estimate the robustness of handling sequencing errors for different OTU picking methods, ten simulated datasets in DBH (Wei and Zhang, 2017) with error rate varies from 0.21 to 0.42% are used to test our DMSC method. Each dataset contains 150,000 sequences from 30 taxa and each taxon contains 5,000 sequences. The OTUs number inferred at 0.05 distance threshold is shown in **Supplementary Figure S6**, from which we can see that with the error rate increase from 0.21 to 0.41%, DMSC infer a smaller number of OTUs than other methods, especially in the 0.33 ∼ 0.41% scope of higher error rate, the OTUs number inferred by DMSC is obviously less than that of other five methods. **Table 4** lists the average OTUs number and std (σ) in the scope of 0.21∼0.41% sequencing errors, from which we can see that the average OTUs number of DMSC is smaller than that of other five methods, and the standard deviation is

TABLE 4 | Average OTUs number and standard deviation of six methods in the scope of 0.21∼0.41% sequencing errors at 0.05 distance threshold.

	DMSC	UCLUST	DBH	CD-HIT	DySC	ESPRIT-Forest
Average OTUs	34	38	37	39	46	92
σ	3.748	9.605	6.863	11.253	3.588	24.691

lower than that of UCLUST, DBH, CD-HIT, and ESPRIT-Forest, near to DySC. **Supplementary Table S7** reports the average OTUs number and std at 0.03 distance threshold, from which we can see that the standard deviation of DMSC is lower than that of other five methods. These results indicate that DMSC can better reduce the OTUs overestimation than the other five methods.

The rapid increase in the amount of sequencing data provides a valuable source to significantly understand bacterial diversity from the environmental samples, meanwhile introducing a serious computational challenge for processing these mass data. In addition to the clustering accuracy, computational complexity is also used to assess a new clustering method. The computational complexity of DMSC mainly contains three components. (1) For generating clusters, a total of N sequences needs to be processed. The large maximum complexity is $O(N)$. (2) In the MCS selection procedure, a distance matrix with size of $\eta \times \eta$ needs to be calculated with a complexity of $O(K \times \eta^2)$, where K is the number of clusters with size larger than η. (3) In the sequences assignment procedure, each sequence is compared with each cluster, resulting in a complexity of $O(K \times N)$. As a result, the total time complexity of DMSC is $O(N+K \times \eta^2+K \times N)$, which is larger than that of traditional heuristic clustering methods such as CD-HIT and UCLUST, but smaller than that of model-based clustering methods such as CROP. In this work, all methods were executed with 16 threads. In order to graphically demonstrate the scaling property of our DMSC method, we compared DMSC with CD-HIT, UCLUST, DBH, DySC, mothur-AL and ESPRIT-Tree on V6 dataset at different sequence size ranging from 1 K to 100 M. **Supplementary Figure S7** shows the running time (wall time) of seven methods. We can see that with the sequence number increases, the speed of DMSC is much faster than mothur-AL, and little lower than the traditional heuristic methods (e.g., CD-HIT, UCLUST, and DBH) that just use one sequence as the seed for each cluster. For the memory usage, **Supplementary Figure S8** graphically describes the memory property of seven methods. From **Supplementary Figure S8**, we can see that DMSC needs a little larger memory usage than the classical greedy clustering methods such as CD-HIT, UCLUST and DySC, and much smaller memory storage than ESPRIT-Forest and mothur-AL for large-scale sequences.

CONCLUSION

16S rRNA high-throughput sequencing has become a powerful and convenient technology for studying microbial diversity and composition in the environmental samples. Until now, numerous heuristic clustering methods have been developed to pick OTUs, but most of them just select one sequence as the cluster seed, resulting in OTUs overestimation and sensitivity to the sequencing errors. In this work, we proposed a novel dynamic multi-seeds heuristic clustering method (namely DMSC) by incorporating the dynamical multi-seeds updating strategy and the heuristic clustering procedure. Meanwhile, DMSC considers the distance's standard deviation within the MCS to generate OTUs. DMSC method is inspired by the idea of seed reselection procedure in DySC, but there are three main differences between DMSC and DySC: (i) DMSC selects MCS as the seeds in one cluster, while DySC just uses one single sequence as the seed; (ii) DySC only updates seed once time, then the seed will be fixed, while DMSC dynamically updates the MCS if a new sequence is added to one cluster, therefore, the seeds is always updated with the cluster size increases; and (iii) a new sequence is assigned to the corresponding cluster depending on the average distance to MCS and the distance standard deviation between each pairwise sequences in MCS, while DySC assigns the new sequence just based on the distance to seed sequence. Compared with the state-of-the-art methods, such as UCLUST, CD-HIT, DBH, DySC, ESPRIT-Forest, CROP, and mothur-AL, the clustering results show that DMSC can produce OTUs with higher quality and reduce OTUs overestimation with low memory usage. Additionally, DMSC is also robust to the sequencing errors.

DATA AVAILABILITY

The DMSC software is available at https://github.com/NWPU-903PR/DMSC, the datasets used and/or analyzed during the current study are available from the corresponding references or from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

Z-GW wrote the code and manuscript and developed the software. S-WZ designed the study and revised the manuscript. Both authors contributed to the conception and design of the study, participated in the data analysis, and to writing and editing of the manuscript. Both authors read, edited, and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (61873202, 61473232, and 91430111).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00428/full#supplementary-material

REFERENCES

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.

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:e00191-16. doi: 10.1128/mSystems.00191-16

Barriuso, J., Valverde, J. R., and Mellado, R. P. (2011). Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. *BMC Bioinformatics* 12:473. doi: 10.1186/1471-2105-12-473

Cai, Y., and Sun, Y. (2011). ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res.* 39:e95. doi: 10.1093/nar/gkr349

Cai, Y., Wei, Z., Jin, Y., Yang, Y., Mai, V., Qi, M., et al. (2017). ESPRIT-Forest: parallel clustering of massive amplicon sequence data in subquadratic time. *PLoS Comput. Biol.* 13:e1005518. doi: 10.1371/journal.pcbi.1005518

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, 335–336.

Chen, S. Y., Deng, F., Huang, Y., Jia, X., Liu, Y. P., and Lai, S. J. (2016). bioOTU: an improved method for simultaneous taxonomic assignments and operational taxonomic units clustering of 16s rRNA gene sequences. *J. Comput. Biol.* 23, 229–238. doi: 10.1089/cmb.2015.0214

Chen, W., Cheng, Y., Zhang, C., Zhang, S., and Zhao, H. (2013a). MSClust: a multi-seeds based clustering algorithm for microbiome profiling using 16S rRNA sequence. *J. Microbiol. Methods* 94, 347–355. doi: 10.1016/j.mimet.2013.07.004

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., and Zhao, H. (2013b). A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS One* 8:e70837. doi: 10.1371/journal.pone.0070837

Cheng, L., Walker, A. W., and Corander, J. (2012). Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Res.* 40, 5240–5249. doi: 10.1093/nar/gks227

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., Mcgarrell, D. M., Sun, Y., et al. (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Edgar, R. (2018). Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ* 6:e5030. doi: 10.7717/peerj.5030

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

Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P. J., Soen, Y., et al. (2018). Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* 6:17. doi: 10.1186/s40168-017-0396-x

Hao, X., Jiang, R., and Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 27, 611–618. doi: 10.1093/bioinformatics/btq725

He, Y., Caporaso, J. G., Jiang, X. T., Sheng, H. F., Huse, S. M., Rideout, J. R., et al. (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* 3:20.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x

Koslicki, D., Foucart, S., and Rosen, G. (2013). Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics* 29, 2096–2102. doi: 10.1093/bioinformatics/btt336

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U.S.A.* 82, 6955–6959.

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.

Liu, Z., Pan, Q., Dezert, J., Han, J.-W., and He, Y. (2018). Classifier fusion with contextual reliability evaluation. *IEEE Trans. Cybern.* 48, 1605–1618. doi: 10.1109/TCYB.2017.2710205

Liu, Z., Pan, Q., Dezert, J., and Martin, A. (2017). Combination of classifiers with optimal weight based on evidential reasoning. *IEEE Trans. Fuzzy Syst.* 26, 1217–1230.

Magoè, 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

Matias Rodrigues, J. F., and von Mering, C. (2013). HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* 30, 287–288. doi: 10.1093/bioinformatics/btt657

Peterson, J., Garges, S., Giovanni, M., Mcinnes, P., Wang, L., Schloss, J. A., et al. (2009). The NIH human microbiome project. *Genome Res.* 19, 2317–2323. doi: 10.1101/gr.096651.109

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584.

Schloss, P. D. (2016). Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. *Msystems* 1:e00027-16.

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. Microbiol.* 77, 3219–3226. doi: 10.1128/AEM.02810-10

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

Schmidt, T. S. B., Matias Rodrigues, J. F., and Mering, C. (2015). Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ. Microbiol.* 17, 1689–1706. doi: 10.1111/1462-2920.12610

Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., et al. (2011). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinform.* 13, 107–121. doi: 10.1093/bib/bbr009

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., Mckendree, W., et al. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res.* 37:e76. doi: 10.1093/nar/gkp285

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* 449, 804–810.

Wang, X., Yao, J., Sun, Y., and Mai, V. (2013). M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinformatics* 14:43. doi: 10.1186/1471-2105-14-43

Wei, Z.-G., and Zhang, S.-W. (2015). MtHc: a motif-based hierarchical method for clustering massive 16S rRNA sequences into OTUs. *Mol. Biosyst.* 11, 1907–1913. doi: 10.1039/c5mb00089k

Wei, Z.-G., and Zhang, S.-W. (2017). DBH: a de Bruijn graph-based heuristic method for clustering large-scale 16S rRNA sequences into OTUs. *J. Theor. Biol.* 425, 80–87. doi: 10.1016/j.jtbi.2017.04.019

Wei, Z.-G., and Zhang, S.-W. (2018). NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. *BMC Bioinformatics* 19:177. doi: 10.1186/s12859-018-2208-0

Wei, Z. G., Zhang, S. W., and Jing, F. (2016). Exploring the interaction patterns among taxa and environments from marine metagenomic data. *Quant. Biol.* 4, 84–91.

Wei, Z. G., Zhang, S. W., and Zhang, Y. Z. (2017). DMclust, a density-based modularity method for accurate OTU picking of 16S rRNA sequences. *Mol. Inform.* 36:1600059. doi: 10.1002/minf.201600059

Westcott, S. L., and Schloss, P. D. (2015). *de novo* clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3:e1487. doi: 10.7717/peerj.1487

Westcott, S. L., and Schloss, P. D. (2017). OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *mSphere* 2:e00073-17. doi: 10.1128/mSphereDirect.00073-17

Zhang, S. W., Wei, Z. G., Zhou, C., Zhang, Y. C., and Zhang, T. H. (2013). "Exploring the interaction patterns in seasonal marine microbial communities with network analysis," in *Proceedings of the International Conference on Systems Biology*, Huangshan, 63–68.

Zheng, Z., Kramer, S., and Schmidt, B. (2012). DySC: software for greedy clustering of 16S rRNA reads. *Bioinformatics* 28, 2182–2183. doi: 10.1093/bioinformatics/bts355

Conflict of Interest Statement: 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.

Copyright © 2019 Wei and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 26 March 2019
doi: 10.3389/fmicb.2019.00507

Identification of Phage Viral Proteins With Hybrid Sequence Features

Xiaoqing Ru[1], *Lihong Li*[1] *and Chunyu Wang*[2]*

[1] School of Information and Electrical Engineering, Hebei University of Engineering, Handan, China, [2] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

The uniqueness of bacteriophages plays an important role in bioinformatics research. In real applications, the function of the bacteriophage virion proteins is the main area of interest. Therefore, it is very important to classify bacteriophage virion proteins and non-phage virion proteins accurately. Extracting comprehensive and effective sequence features from proteins plays a vital role in protein classification. In order to more fully represent protein information, this paper is more comprehensive and effective by combining the features extracted by the feature information representation algorithm based on sequence information (CCPA) and the feature representation algorithm based on sequence and structure information. After extracting features, the Max-Relevance-Max-Distance (MRMD) algorithm is used to select the optimal feature set with the strongest correlation between class labels and low redundancy between features. Given the randomness of the samples selected by the random forest classification algorithm and the randomness features for producing each node variable, a random forest method is employed to perform 10-fold cross-validation on the bacteriophage protein classification. The accuracy of this model is as high as 93.5% in the classification of phage proteins in this study. This study also found that, among the eight physicochemical properties considered, the charge property has the greatest impact on the classification of bacteriophage proteins These results indicate that the model discussed in this paper is an important tool in bacteriophage protein research.

Keywords: phage virion proteins, machine learning, feature extraction, feature selection, hybrid sequence features

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Zhiwei Ji,
University of Texas Health Science Center, United States
Nuria Quiles Puchalt,
University of Glasgow, United Kingdom

***Correspondence:**
Chunyu Wang
chunyu@hit.edu.cn

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

Received: *24 December 2018*
Accepted: *27 February 2019*
Published: *26 March 2019*

Citation:
Ru X, Li L and Wang C (2019) Identification of Phage Viral Proteins With Hybrid Sequence Features. Front. Microbiol. 10:507. doi: 10.3389/fmicb.2019.00507

INTRODUCTION

In the biological world, bacteriophages are ubiquitous, with different genomes and lifestyles. According to their morphology, they can be classified as either tail, tailless, or filamentous bacteriophages. According to morphology and nucleic acid, phages are classified as infect bacteria and infect archaea. The bacteriophage must be attached to a host cell for growth and reproduction (Seguritan et al., 2012), and directly affects the host population by lysing host cells. In addition, each bacteriophage is specific and greatly reduces the damage to host cells (Haq et al., 2012). Identification and classification of various bacteria can be performed based on the universality, diversity, dependence, and specificity of bacteriophages (Marks and Sharp, 2015).The structure of bacteriophages is simple, consisting of only a protein shell and genetic material (DNA or RNA) (Haq et al., 2012), making them important substances for simplifying experimental research in bioinformatics. As a bacteriophage can insert genes into host cells (Ding et al., 2014), it is an important tool for studying genetics (Cheng et al., 2018; Hu et al., 2018). Hershey

(Hershey and Chase, 1952) performed biological experiments using the T2 bacteriophage and bacteria in 1952, and finally confirmed that DNA is the genetic material of bacteriophages and other organisms. The significance of this research in the development of biological science earned Hershey and coworkers the Nobel Prize in Physiology. Bacteriophage provide experimental systems and tools for the molecular biological science revolution. The bacteriophage rapid development has led to dection of basic principles of ecology and evolution. Besides, it is relatively easy to synthesize and has modular characteristic, which cater to the needs of synthetic biologists and carry out engineering research and implementation of biological function.

Bacteriophage proteins are classified into virion and non-viron proteins (Zhang et al., 2015), with most practical interest focusing on the function of bacteriophage virion proteins (Feng et al., 2013b). Therefore, bacteriophage proteins must be accurately classified and identified so that researchers can further study the structure and function of a particular bacteriophage. After the human genome project was officially launched in 1990, the number of bacteriophage protein sequences with unknown functions increased dramatically (Seguritan et al., 2012; Chen et al., 2018a). Faced with a large volume of data, traditional biological experimental methods could no longer keep up with the post-gene era (Chen W. et al., 2016; Cheng et al., 2019; Mrozek et al., 2016; Hu et al., 2018). For this reason, researchers introduced different machine learning algorithms into bacteriophage classification and prediction research. For example, Li et al. (2007) developed a support vector machine system called SynFPS that uses the gene–gene distance determined by k-means clustering to identify closely related genomes and perform gene function prediction. Using the protein appearance frequency of amino acids and information of isoelectric points, Seguritan et al. (2012) developed an artificial neural network method to classify viral structures. Feng et al. (2013b) used the main amino acid and dipeptide components as an encoding scheme, and modified a naive Bayes classifier to identify bacteriophage proteins. Ding et al. (2014) used g-gap dipeptide composition to represent protein sequence information, incremental feature selection to analyze the variance and identify the optimal feature set, and a support vector machine for classification. Zhang et al. (2015) obtained sequence feature vectors with various techniques, and then used the incremental feature selection algorithm to select the optimal feature subsets. Finally, the prediction results of individual classifiers trained in different feature spaces were integrated to produce the final classification effect. Machine learning algorithm (Robert, 2012; Stephenson et al., 2018) automatically analyze and obtain rules from data and use them to predict unknown data (Chen and Yan, 2013; Yu et al., 2015, 2016a; Chen and Huang, 2017; Chen et al., 2018h; Wang et al., 2018). This saves time and money, but the results from such algorithms are not as convincing as those from biological experiments. Therefore, it is especially important to choose an appropriate machine learning algorithm to ensure the most accurate classification results (Liu, 2017; Yao et al., 2017; Yu et al., 2017a). In a protein classification experiment, the classification effect depends largely on the feature set extracted (Zou et al., 2013; Bin et al., 2015; Mrozek et al., 2015; Jia et al., 2016; Yu et al., 2016b, 2018; Zhang et al., 2016; Huang et al., 2017; Qu et al., 2017; Jiang et al., 2018; Qiao et al., 2018; Xiong et al., 2018; Xu et al., 2018b). To date, feature extraction methods are divided into sequence-based and structure-based approaches (Huang et al., 2017; Qu et al., 2017) The feature set extraction part of this study is obtained by combining the features extracted by the two feature extraction methods.

In this study, we examined the final classification effect of the selected methods and the stability of the dataset when the feature dimension was reduced. First, to remove the imbalance in the reference dataset, CD-Hit was used to remove redundant data, resulting in a balanced dataset that contains comprehensive information and less redundancy. Pearson's correlation coefficient and three distance functions (Euclidean and cosine distances and the Tanimoto coefficient) (Zou et al., 2016) were then used to calculate the correlation between features and class labels and the redundancy between features. Finally, the optimal feature subset with the strongest correlation between features and class labels and low redundancy between features was selected. According to some recent studies(Wu et al., 2009; Yi et al., 2011; Chen and Lin, 2012; Yang et al., 2015; Yu et al., 2017b; Zhang and Liu, 2017; Xu et al., 2018a; Liu et al., 2019), the best algorithms for protein classification are support vector machines and random forest algorithms. However, support vector machines are more suitable for small sample sets in which the number of dimensions is greater than the number of samples. Thus, the random forest algorithm was used in this study. The random forest algorithm (Breiman, 2001; Yao et al., 2017) combines multiple weak classifiers to produce a final result that has higher accuracy and better generalization performance. It can achieve good results, mainly because of the random nature of the "forest," which makes the algorithm resistant to overfitting and more precise. Finally, in terms of bacteriophage protein classification, the data set extracted by combining the features and the feature selection of the feature set have a positive impact on the protein classification effect. Our results also show that, among the eight physicochemical properties of amino acids, the charge property has the greatest influence on the classification of bacteriophage proteins. To evaluate the performance of the models used in this study, the results were compared with those given by the methods introduced in (Feng et al., 2013b; Ding et al., 2014; Zhang et al., 2015). **Figure 1** shows the workflow of this study.

METHODS

Dataset Processing

Source: UniProt (Rolf, 2004; Consortium, 2012) is a widely used protein sequence database that offers low protein sequence redundancy and complete protein function interpretation (Cao and Cheng, 2016a; Jiang et al., 2016). As this website is free and open, researchers can download the desired protein sequence for free. The original positive samples used in this study (a total of 15,765 data), e.g., the number of bacteriophage virion proteins, were downloaded from this database. After obtaining the bacteriophage virion protein (positive) sample set, the PFAM family of positive samples was excluded from all PFAM families,

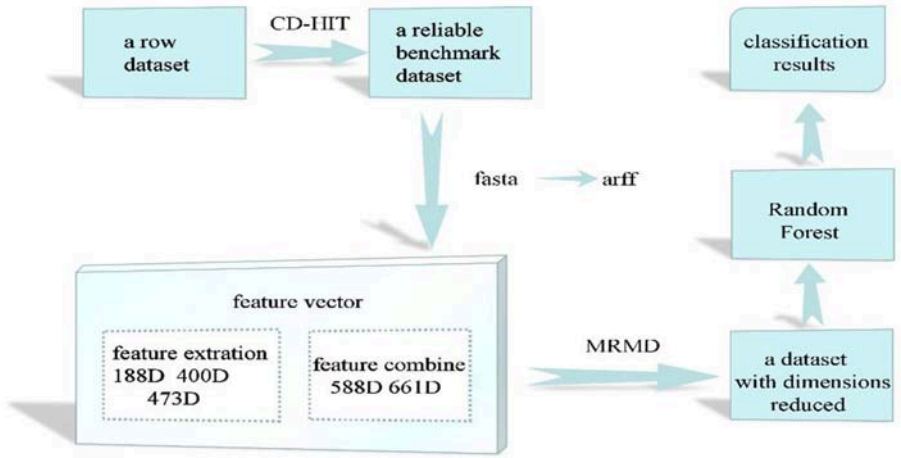


FIGURE 1 | Outline flowchart of this study.

such that the remaining samples were families of non-phage virion proteins. Finally, the longest protein sequence of the remaining families was extracted to form a negative sample set. The positive and counterexample datasets obtained as described above may all contain homologous sequences. Using such sample sets would result in the classification accuracy being overestimated, which is not conducive to the establishment of prediction models. Therefore, we used the CD-Hit tool to remove redundant positive and negative samples from the datasets.

Data integration: The CD-Hit (Li et al., 2001; Li and Godzik, 2006; Huang et al., 2010; Fu et al., 2012; Chen et al., 2017) redundancy tool effectively clusters similar sequences. The basic principle is to sort protein sequences in the dataset in descending order. The longest sequence is taken as the first class, and then this is compared with the second-longest protein sequence in terms of their similarity. If the similarity between the two is greater than some threshold, they are deemed to belong to the same class. Otherwise, the second-longest sequence forms a new class. Because the bacteriophage virion protein sequences were downloaded from UniProt, which ensures relatively low redundancy, the interrupt threshold was set to 0.8. The non-phage virion proteins had a higher degree of redundancy, so their interrupt threshold was set to 0.4. Thus, 6,251 bacteriophage virion protein sequences and 9,514 non-phage virion protein sequences were obtained. The union of the resulting positive and negative sample datasets gives the total dataset, and the intersection of the two is empty.

Feature Extraction

Representation Algorithms for Amino Acid Composition and Eight Physicochemical Properties

In this study, a feature set containing 188 dimensions was extracted based on amino acid composition and eight physicochemical properties. The amino acid composition is one of the most basic features of proteins (Zhang et al., 2015; Cao and Cheng, 2016b). Eight physicochemical properties of amino acids also play a role in the functional properties of bacteriophage proteins. In 1988, Coia et al. (1988) found that amino acids having lighter side chain groups are more likely to constitute bacteriophage virion sequences. In 1994, Marvin et al. (1994) proposed that hydrophilicity, hydrophobicity, and charge have a greater impact on the function of bacteriophage virion proteins. In 2008, Shen and Chou (2008) identified the vital role that the hydrophilicity and hydrophobicity of amino acids play in the folding of proteins. In 2014, Ting et al. (2014) used logistic regression to integrate several biological features, including physicochemical properties for predicting lysine acetylation, thus demonstrating the effect of physicochemical properties on protein structure and function. Therefore, the amino acid composition and its eight physicochemical properties are used to extract features that reflect the characteristics of bacteriophage proteins.

The 20 most common amino acids are as follows:

$$CAA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} \quad (1)$$

The occurrence frequency of each amino acid in a protein sequence can be expressed as:

$$f_{1i} = \left\{ \frac{n_i}{L} | 1 \le i \le 20 \right\} \quad (2)$$

Where n_i is the frequency with which amino acid i occurs in the protein sequence and L is the length of the protein sequence.

In addition, these 20 amino acids can be classified into three types according to their physicochemical properties (Chou and Com, 2010), as shown in **Figure 2**.

The composition, transformation, and distribution of amino acids were determined by Dubchak et al. (1995) based on a global description of protein sequences. The feature extraction methods for the eight physicochemical properties of a protein sequence are as follows. Taking the electrode polarity as an example

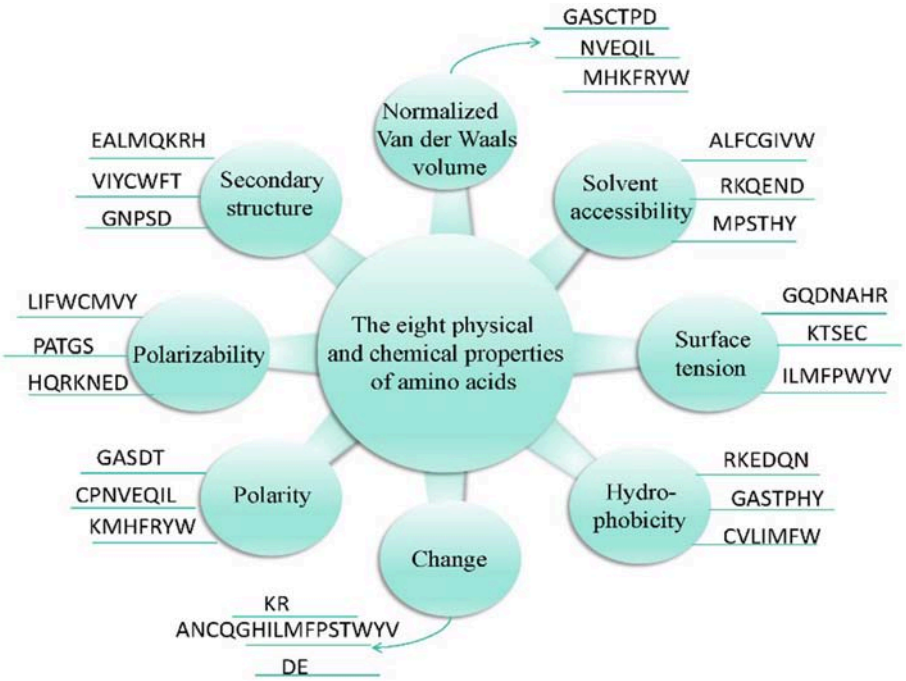


FIGURE 2 | Eight physicochemical properties of amino acids.

(expressed by p), the 20 amino acids are divided into high-, medium-, and low-charged polarity groups, which are expressed by p_h, p_p, p_l, respectively. The composition, transformation, and distribution of the amino acids at this time can be represented by equations (3)–(7).

Composition features (Dubchak et al., 1995) (frequency of each charged electrode group in a sequence):

$$\left(f_{21}, f_{22}, f_{23}\right) = \left[\frac{n_1 p_h}{L}, \frac{n_2 p_p}{L}, \frac{n_3 p_l}{L}\right] \tag{3}$$

where f_{21}, f_{22}, f_{23} denote the content of the high-, medium-, and low-charged polarity groups in a sequence, respectively,L is the length of the protein sequence,n_1, n_2, n_3 are the frequencies with which the three electrode groups appear in the sequence.

Conversion feature (Dubchak et al., 1995) (frequency of occurrence of bigeminal sequences):

$$\left(f_{31}, f_{32}, f_{33}\right) = \left[\frac{m_1 p_{hl}}{L-1}, \frac{m_2 p_{hp}}{L-1}, \frac{m_3 p_{pl}}{L-1}\right] \tag{4}$$

Where f_{31}, f_{32}, f_{33} denote the content of the three bigeminal groups p_{hl}, p_{hp}, p_{pl}, and m_1, m_2, m_3 are the frequencies of these three bigeminal groups appearing in sequence. There are three possible sequences of the charged polarity: p_{hl}, p_{hp}, p_{pl} In addition, in a protein sequence of length L, assuming that any two adjacent amino acids constitute a pair, the protein sequence contains $L-1$ paired sequences (Zou et al., 2013).

Distribution features (Dubchak et al., 1995) (amino acid distribution of the high-, medium-, and low-charged polarity groups):

$$\left(f_{411}, f_{412}, f_{413}, f_{414}, f_{415}\right)^T = \left[a_{1\%}, a_{25\%}, a_{50\%}, a_{75\%}, a_{100\%}\right]^T \tag{5}$$

$$\left(f_{421}, f_{422}, f_{423}, f_{424}, f_{425}\right)^T = \left[b_{1\%}, b_{25\%}, b_{50\%}, b_{75\%}, b_{100\%}\right]^T \tag{6}$$

$$\left(f_{431}, f_{432}, f_{433}, f_{434}, f_{435}\right)^T = \left[c_{1\%}, c_{25\%}, c_{50\%}, c_{75\%}, c_{100\%}\right]^T \tag{7}$$

Where $a_{1\%}, a_{25\%}, a_{50\%} a_{75\%} a_{100\%}$ represent the positions of the first, 25, 50, 75, and 100% high-charged polarity groups in a sequence, $b_{1\%}, b_{25\%}, b_{50\%}, b_{75\%}, b_{100\%}$ represent the positions of the first, 25, 50, 75, and 100% medium-charged polarity groups in a sequence and $c_{1\%}, c_{25\%}, c_{50\%}, c_{75\%}, c_{100\%}$ represent the positions of the first, 25, 50, 75, and 100% low-charged polarity groups in a sequence.

In summary, $(3 + 3 + 3 \times 5) = 21$-dimensional features can be extracted from each physicochemical property, and so $8 \times 21 = 168$-dimensional features can be extracted from the eight physicochemical properties. The 188-dimensional features (20-dimensional + 168-dimensional) are used to express the characteristics of bacteriophage proteins, and are extracted based on the content ratio of each of the 20 amino acids in the sequence and the eight physicochemical properties.

Adaptive k-skip-*n*-Gram Algorithm

A feature set containing 400 dimensions is extracted based on the adaptive k-skip-n-gram method (Feng et al., 2013c; Cao et al., 2017; Wei et al., 2017a; Tang et al., 2018) . In this study, the value of n was set to 2 ($20^2 = 400$).

The K value represents the separation distance between two amino acids. For example, in the protein sequence $S =$

$A_1A_2A_3 \cdots A_L$ (where L is the length of the sequence),

$$K = i - j - 1 \tag{8}$$

And A_i,A_j are the ith and jth amino acids of S.

In a bacteriophage protein dataset, the sequences have very different lengths. If the parameter K is fixed to a specific value, the sequence information cannot be properly represented, which will affect the final classification effect. Therefore, the value of k was set to be adaptive so that K could vary with the length of the sequence.

For $n = 2$, the combinations of the 20 most common amino acids and the number of occurrences of each combination in the sample datasets are as shown in **Figure 3**.

This process is similar to full connection in a neural network. Among the 20 common amino acids, anyone can combine with another amino acid (or itself) in pairs, and the combination is random. In the same way as full connection, this leads to overfitting when there are too many data. Therefore, n should not be too high when using an adaptive k-skip-n-gram method. When $n = 1$, we have the traditional n-gram model proposed by Guthrie et al. (2006), which does not apply to shorter protein sequences. Therefore, n was set to 2 in this study.

In this feature extraction method, the combination set of two specified interval amino acids (Wei et al., 2017a) is given by:

$$\begin{cases} skip\,(K=0) = \{A_1A_2, A_2A_3, \cdots, A_{L-1}A_L\} \\ skip\,(K=1) = \{A_1A_3, A_2A_4, \cdots, A_{L-2}A_L\} \\ \qquad\qquad \vdots \\ skip\,(K=k) = \{A_1A_{2+k}, A_2A_{3+k}, \cdots, A_{L-k+1}A_L\} \end{cases} \tag{9}$$

In addition, C is used to represent a set of two amino acids that are combined at all intervals in a sequence (Wei et al., 2017a).Namely:

$$C_{skipgram} = \left\{ \bigcup\nolimits_{d=0}^{k} skip(K=d) | d = 1, 2, 3, \cdots k \right\} \tag{10}$$

Finally, the feature extraction formula (Wei et al., 2017a) is:

$$FV = \left\{ \frac{N(a_{m1}a_{m2}\cdots a_{mn})}{N(C_{skipgram})} | 1 \le m_i \le 20, 1 \le i \le n \right\} \tag{11}$$

Where $N(C_{skipgram})$ is the total number of elements in set C,$a_{m1}a_{m2}\cdots a_{mn}$ are the 20^n kinds of amino acid combinations of length n, $N(a_{m1}a_{m2}\cdots a_{mn})$ is the frequency that the two-two combination in $a_{m1}a_{m2}\cdots a_{mn}$ occurs in $C_{skipgram}$

Mixed Representation Algorithm (Seq-Str)

Some researchers have combined different feature extraction methods and achieved very good classification results (Dehzangi et al., 2013; Zou et al., 2014; Leyi et al., 2015, 2018; Chen X. et al., 2016; Ding et al., 2016, 2017a,b; Li et al., 2016; Chen et al., 2017,a,b, 2018c,d,e; Su et al., 2018 Shen et al., 2019; Wei et al., 2019; Zhu et al., 2019). Wei et al. (2015) proposed a novel feature extraction method that uses both the profile of PSI-BLAST (Altschul et al., 1997) and the profile of PSI-PRED (Jones, 1999), which contain rich evolutionary information and secondary structure information, respectively. In this way, the 473-dimensional feature can be extracted.

1) Extract 20-dimensional features based on PSI-BLAST as follows:

$$\mathrm{FV} = \left\{ \overline{S}_i = \frac{1}{L} \sum\nolimits_{z=1}^{L} S_{z,i} |\ i = 1, 2, \ldots 20 \right\} \tag{12}$$

$S_{z,i}$indicates that during the evolution process, the residue at the "z" position in the sequence S is mutated to the fraction of the "i" species, and "i" is one of the 20 common residues. S_i indicates that during the evolution, the residue in sequence S is mutated to the average score of the ith residue.

2) Extracting 420-dimensional features based on n-gram: The Adaptive k-skip-n-gram algorithm that does not consider the k value is the n-gram method. Here, take n equal to 1 and n equal to 2

3) Based on the secondary structure sequence, the following six features are extracted (Wei et al., 2015): Three feature extraction formulas for spatial arrangement

$$CMV_H = \sum\nolimits_{z=1}^{n_H} P_{H_z}/L(L-1) \tag{13}$$

Where P_{H_z}represents the position index of the zth H in the secondary structure of the sequence S. n_H represents the total number of occurrences of H in the secondary structure of sequence.

Two feature extraction formulas for the percentage of the maximum continuous length (Wei et al., 2015).

$$Rmax_{C_H} = max\{C_H\}/L \tag{14}$$

C_H represents the length of the fragment in which H appears consecutively in the sequence of the secondary structure.

A new feature for distinguishing between two structural classes, $\alpha + \beta$ and $\frac{\alpha}{\beta}$: (Wei et al., 2015)

$$f_{\beta\alpha\beta} = n_{\beta\alpha\beta}/L_{seg} - 2 \tag{15}$$

This formula calculates the frequency at which βαβ appears in the fragmented sequence S_{seg}, $n_{\beta\alpha\beta}$ represents the number of times βαβ appears in S_{seg}, L_{seg}indicates the length of S_{seg}.

4) Extracting 27 features based on structural probability matrices: Three features from the overall information and 24 features from local information

Feature Selection

Based on the feature extraction methods described in section Feature extraction, We extracted a 188-dimensional, 400-dimensional feature set based on sequence information, and a 473-dimensional data set based on sequence and secondary structure information representing the entire bacteriophage protein sequence dataset. Some redundant or irrelevant cases

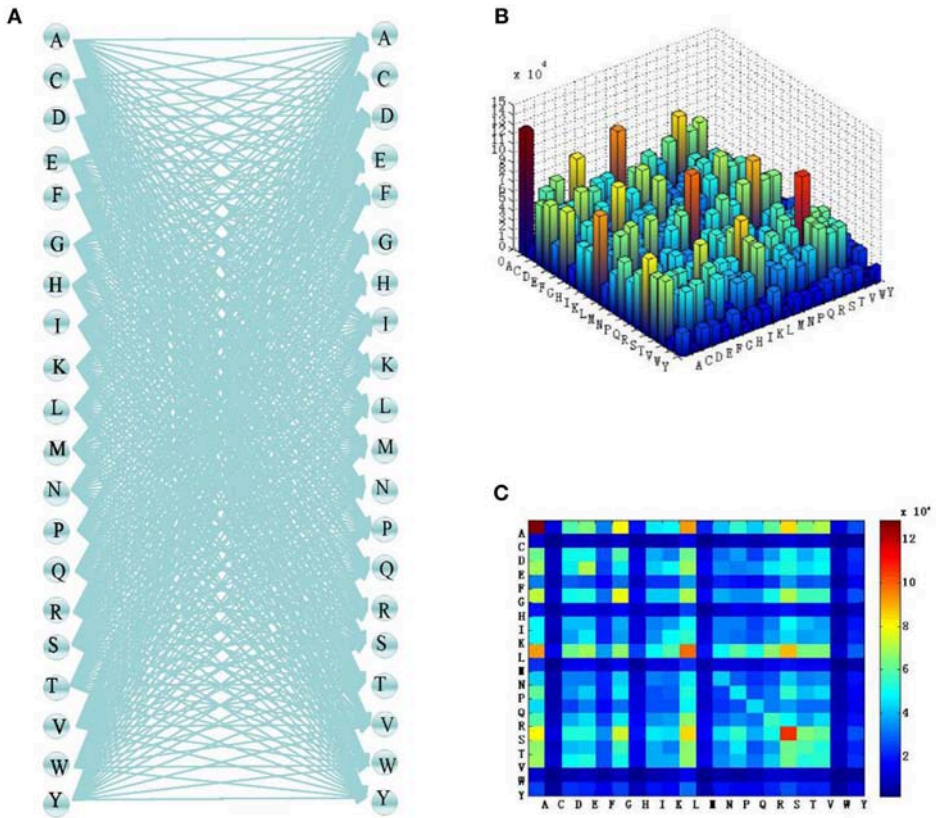


FIGURE 3 | Two-two combination process of amino acids. **(A)** Two-two combination of residues. **(B)** Three-dimensional heat map of amino acid frequency. **(C)** Heat map of amino acid frequency.

TABLE 1 | Classification results of three data sets under different classification algorithms.

Feature_ extraction	Feature_ selection	number of D	LibSVM (%)	Naive Bayes (%)	Random forest (%)
CCPA		188D	68.5	78.3	91.3
	MRMD	185D	68.5	78.3	91.5
AKSNG		400D	60.3	71.8	88.7
	MRMD	252D	60.3	72.8	89.0
Seq-Str		473D	80.6	80.9	92.6
	MRMD	189D	82.0	83.1	93.2

TABLE 2 | Classification performance under different feature extraction methods.

Extraction method	Number of D	SN (%)	SP (%)	ACC (%)	MCC (%)
Seq based	188D	87.4	93.6	91.3	81.5
	400D	82.8	92.4	88.7	76.1
Seq and str based	473D	86.2	97.2	92.6	85.1
Com based	588D	87.1	93.2	91.2	80.7
	661D	87.5	96.5	93.1	85.3

were still present in these features. The existence of invalid features wastes time and computational resources, and affects the classification accuracy of the model (Chen et al., 2018b,f,g; Dao et al., 2018; Yang et al., 2018; Zhu et al., 2018a,b). In this paper, the Max-Relevance-Max-Distance (MRMD) (Zou et al., 2016) method was used to select features and identify higher-quality feature sets, i.e., the optimal feature subset. In this method, Pearson's correlation coefficient is used to calculate the correlation between features and class labels (*MR*), thus enabling the selection of features with strong correlation to the target class. Three distance functions (Euclidean and cosine distances and the Tanimoto coefficient) are used to calculate the redundancy between features (*MD*) and identify features with low redundancy.

Taking the two eigenvectors (X,Y) as an example, Pearson's correlation coefficient (Pearson, 1909) expressed as follows:

$$\rho_{X,Y} = corr\,(X,Y) = \frac{cov\,(X,Y)}{\sigma_X \sigma_Y} \tag{16}$$

Where σ_X and σ_Y denote the standard deviation of the two vectors, $cov(X,Y)$ is the covariance, which is used to measure the relationship between two random variables. The covariance

formula is as follows:

$$cov\,(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1} \tag{17}$$

Where $\bar{X}$ and $\bar{Y}$ denote the mean of the respective vectors.

The formula for the Euclidean distance (Larson and Edwards, 1991; Deza and Deza, 2009) is:

$$ED_i = \frac{1}{M-1}\sum\sqrt{\sum_{q=1}^{n}(x_q - y_q)^2} \tag{18}$$

Where M is the number of feature vectors,n is the total number of elements in each vector, and x_q, y_q are the q-th elements in X, Y, respectively.

The cosine distance formula (Tan et al., 2005) is:

$$COS_i = \frac{1}{M-1}\sum\left(\frac{X\cdot Y}{||X||\cdot||Y||}\right) \tag{19}$$

Where

$$\|X\| = \sqrt{\sum_{q=1}^{n} {x_q}^2} \tag{20}$$

The Tanimoto coefficient (Rogers and Tanimoto, 1960) is given by:

$$TC_i = \frac{1}{M-1}\sum\left(\frac{X\cdot Y}{||X||^2 + ||Y||^2 - X\cdot Y}\right) \tag{21}$$

Using these distance metrics, we identified the features with the strongest correlation and minimum redundancy with respect to the class labels. In different scenarios, we can increase the weights of MR and MD $\left(max\left(wr\times MR_i + wd\times MD_i\right)\right)$ to ensure the acquired features are suitable for the classification task.

EXPERIMENTS

Performance Evaluation Criteria

A 10-fold cross-validation method was employed to evaluate the models. There are four common evaluation indicators, namely the accuracy (ACC), sensitivity (SN), specificity (SP), and Matthews' correlation coefficient (MCC) (Feng et al., 2013a, 2018; Chen W. et al., 2016; Wei et al., 2017b,c; Xu et al., 2017; Jingjing et al., 2018). These are expressed as follows (Zou et al., 2013; Chen et al., 2014; Qu et al., 2017):

$$SN = \frac{TP}{TP+FN} \tag{22}$$

$$SP = \frac{TN}{TN+FP} \tag{23}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{24}$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FN)\,(TP+FP)\,(TN+FP)\,(TN+FN)}} \tag{25}$$

Where TP denotes true positive, i.e., the number of positive samples that are predicted to be positive samples, TN denotes true negative, i.e., the number of negative samples that are predicted to be negative samples, FP denotes false positive, i.e., the number of negative samples that are predicted to be positive samples, and FN denotes false negative, i.e., the number of positive samples that are predicted to be negative samples.

Classification Effects of Different Classifiers

Experiment 1: This part of the experiment is based on the feature sets of 188, 400, and 473 dimensions extracted by the method in Feature extraction. The accuracy of each classification algorithm before and after using the MRMD feature selection algorithm is presented in **Table 1**.

The data in **Table 1** indicate that, for the classification of bacteriophage proteins, no matter which feature extraction algorithm is used, whether or not feature selection is performed, the random forest algorithm is the best classification effect.

Performance of Different Feature Extraction Methods

Experiment 2: Experiment 1 showed that the random forest algorithm produces the best classification of bacteriophage proteins. In this second experiment, the 188-dimensional and 400-dimensional datasets extracted based on sequence information (Seq Based), a 473-dimensional dataset extracted based on structure (Seq and stru Based), and two combined feature sets (Com Based) were integrated into the random forest algorithm, and the resulting performance was compared. The experimental results are presented in **Table 2**.

TABLE 3 | Classification performance under each model.

Model	Feature_extraction	SN (%)	SP (%)	ACC (%)	MCC (%)
Mode l	CCPA (188)	87.5	93.4	91.5	81.4
Mode 2	AKSNG (400)	82.9	92.2	89.0	76.0
Mode 3	Seq-Str (473)	86.7	96.6	93.2	84.8
Mode 4	Combine (588)	87.6	93.5	91.5	81.5
Mode 5	Combine (661)	87.9	96.3	93.5	85.3

TABLE 4 | Performance comparison against recent methods.

Model	SN (%)	SP (%)	ACC (%)	MCC (%)
Feng et al. (2013b)	75.7	80.7	79.1	54.9
Ding et al. (2014)	75.7	89.4	85.0	65.5
Zhang et al. (2015)	87.0	83.0	85.0	70.1
This search	87.9	96.3	93.5	85.3

TABLE 5 | Impact of physicochemical properties on classification.

NO.	Fea name	Score	Implication
1	Fea 120	1.0	Position of the 100%th neutral electrical storage amino acid in a sequence
2	Fea 157	0.9968696407744475	Position of the 100%th helical amino acid in a sequence
3	Fea 178	0.9950260206126923	Position of the 100%th soluble amino acid in a sequence
4	Fea 99	0.9949600329187752	Position of the 100%th neutral polarizability amino acid in a sequence
5	Fea 136	0.9948079966447566	Position of the 100%th large tensile amino acid in a sequence
6	Fea 83	0.994509178771573	Position of the 100%th high-electrode amino acid in a sequence
7	Fea 52	0.994137797849692	Position of the 100%th small van der Waals volume amino acid in a sequence
8	Fea 31	0.9937317569946658	Position of the 100%th hydrophilic amino acid in a sequence

Feature fusion can boost the recognition performance by combining the complementary information of different features (Zhu et al., 2016, 2018c). A 588-dimensional feature set was obtained by combining the features of the 188- and 400-dimensional feature sets, and a 661-dimensional feature set was obtained by combining the features of the 188- and 473-dimensional feature sets. According to the experimental results, the 188-, 473-, 588-, and 661-dimensional feature set models give better bacteriophage protein classification performance, However, based on the data of the other three evaluation indicators, the 661-dimensional feature set obtained by combining the 188-dimensional feature set extracted based on the sequence information and the features of the 473-dimensional feature set extracted based on the sequence and the secondary structure is the best. This indicates that the feature set extracted by the feature representation algorithm containing both sequence information and structural information in phage protein classification has the best influence on the classification effect, and also shows that combining some feature sets in protein classification is effective for improving classification performance.

Importance of Feature Selection

Experiment 3: This experiment used the random forest classification algorithm to classify the feature sets after MRMD. The results are given in **Table 3**.

The comparison of the data in **Tables 2**, **3** shows that after using the feature selection algorithm (MRMD), the classification effect does not change with the decrease of the dimension, and even with the decrease of the dimension, the classification effect becomes better. After removing the redundant features, the best classification performance is still the data set obtained by feature combination, that is, the 256-dimensional feature set obtained by removing redundant features from the 661-dimensional feature set.

Comparison With Recent Methods

Experiment 4: To provide an objective demonstration of the performance of the model described in this paper, this experiment compared the optimal proposed model with bacteriophage protein classification models proposed in recent years. The results are presented in **Table 4**.

It is clear from **Table 4** that the bacteriophage classification model proposed in this paper achieves a good classification effect, with a classification accuracy of 93.5%. Compared with Feng, it has increased by 14%, compared with Ding and Zhang by 8%. In the other three evaluation indicators, there are also different degrees of improvement, indicating that the model proposed in this paper is an effective tool for phage protein classification.

Analyzing the Impact of Eight Physicochemical Properties

This section summarizes the first eight dimensional features that have a significant impact on the classification effect of bacteriophage proteins. The top eight features are listed in **Table 5** in order of their impact.

According to the information in this table, the effects of eight physicochemical properties of amino acids on the classification of bacteriophage proteins are evenly distributed, and that which has the greatest impact on the classification is the charge property of amino acids.

CONCLUSION

Bacteriophage proteins are of special significance for cell typing and pathological research. It is very important to correctly classify virion and non-virion bacteriophage proteins. Therefore, this paper has proposed the following classification model: (1) higher-quality feature datasets are extracted with extraction algorithms based on feature combination; (2) the optimal feature subset is selected using the MRMD algorithm for feature selection; and (3) the random forest algorithm is applied to perform protein classification. The model can achieve accuracy of up to 93.5% for the classification of bacteriophage proteins. This demonstrates that the model developed in this paper is an important tool for the classification of bacteriophage proteins. For the future direction, link prediction paradigms, which have been successfully applied in the prediction of disease genes (Zeng et al., 2017) and miRNAs (Liu et al., 2016; Zeng et al., 2018), can be considered for identification of bacteriophage proteins. It might also be important to integrate evolutionary information using tools like evolutionary trees and networks (Yang et al., 2013, 2014). Finally, computational intelligence such as neural networks (Song et al., 2018a,b) and evolutionary algorithms (Hang et al., 2018) can be applied in this field.

AUTHOR CONTRIBUTIONS

XR implemented the experiments and drafted the manuscript. LL and CW initiated the idea, conceived the whole process, and finalized the paper. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The work was supported Natural Science Foundation of China (No.61872114, 91735306), and the National Key Research and Development Plan Task of China (No. 2016YFC0901902). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Bin, L., Fule, L., Xiaolong, W., Junjie, C., Longyun, F., and Kuo-Chen, C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cao, R., and Cheng, J. (2016a). Integrated protein function prediction by mining function associations, sequences, and protein-protein and gene-gene interaction networks. *Methods* 93, 84–91. doi: 10.1016/j.ymeth.2015.09.011

Cao, R., and Cheng, J. (2016b). Protein single-model quality assessment by feature-based probability density functions. *Sci. Rep.* 6:23990. doi: 10.1038/srep23990

Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., and Chen, Z. (2017). ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 22:1732. doi: 10.3390/molecules22101732

Chen, J., Guo, M., Wang, X., and Liu, B. (2018a). A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief Bioinform.* 19, 231–244. doi: 10.1093/bib/bbw108

Chen, W., Feng, P., Ding, H., and Lin, H. (2018b). Classifying included and excluded exons in exon skipping event using histone modifications. *Front. Genet.* 9:433. doi: 10.3389/fgene.2018.00433

Chen, W., Feng, P., Tang, H., Ding, H., and Lin, H. (2016). RAMPred: identifying the N1-methyladenosine sites in eukaryotic transcriptomes. *Sci. Rep.* 6:31080. doi: 10.1038/srep31080

Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K. C. (2018c). iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. Molecular therapy. *Nucleic Acids* 11, 468–474. doi: 10.1016/j.omtn.2018.03.012

Chen, W., Feng, P. M., Lin, H., and Chou, K. C. (2014). iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed Res. Int.* 2014, 1–12. doi: 10.1155/2014/623149

Chen, W., and Lin, H. (2012). Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. *Comput. Biol. Med.* 42, 504–507. doi: 10.1016/j.compbiomed.2012.01.003

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Chen, X., Guan, N. N., Sun, Y. Z., Li, J. Q., and Qu, J. (2018d). MicroRNA-small molecule association identification: from experimental results to computational models. *Brief. Bioinform.* 2018:bby098. doi: 10.1093/bib/bby098

Chen, X., and Huang, L. (2017). LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Sun, Y. Z., Guan, N. N., Qu, J., Huang, Z. A., Zhu, Z. X., et al. (2018e). Computational models for lncRNA function prediction and functional similarity calculation. *Brief Funct. Genomics* 18, 58–82. doi: 10.1093/bfgp/ely031

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018f). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018g). BNPMDA: bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2017b). *MicroRNAs and complex* diseases: from experimental results to computational models. *Brief. Bioinform.* 2017:bbx130. doi: 10.1093/bib/bbx130

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017a). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016). Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chen, X., Yin, J., Qu, J., and Huang, L. (2018h). MDHGI: Matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 34820. doi: 10.1038/srep34820

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief. Bioinform.* 20, 203–209. doi: 10.1093/bib/bbx103

Chou, K., and Com, M. P. (2010). Prediction of protein cellular attributes using pseudo-amino acid composition. *Protein Struct. Funct. Bioinform.* 43:246–255. doi: 10.1002/prot.1035

Coia, G., Parker, M. D., Speight, G., Byrne, M. E., and Westaway, E. G. (1988). Nucleotide and complete amino acid sequences of Kunjin virus: definitive gene order and characteristics of the virus-specified proteins. *J. Gen. Virol.* 69, 1–21. doi: 10.1099/0022-1317-69-1-1

Consortium, U. P. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 40, D71–D75. doi: 10.1093/nar/gkr981

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2018). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics*. 2018:bty943. doi: 10.1093/bioinformatics/bty943

Dehzangi, A., Paliwal, K., Sharma, A., Dehzangi, O., and Sattar, A. (2013). A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 564–575. doi: 10.1109/TCBB.2013.65

Deza, M. M., and Deza, E. (2009). Encyclopedia of distances. *Refer. Rev.* 24, 1–583. doi: 10.1007/978-3-642-00234-2

Ding, H., Feng, P. M., Chen, W., and Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection

and analysis. *Mol. Biosyst.* 10, 2229–2235. doi: 10.1039/C4MB00316K

Ding, Y., Tang, J., and Guo, F. (2016). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinform.* 17:398. doi: 10.1186/s12859-016-1253-9

Ding, Y., Tang, J., and Guo, F. (2017a). Identification of protein-ligand binding sites by sequence information and ensemble classifier. *J. Chem. Inf. Model.* 57, 3149–3161. doi: 10.1021/acs.jcim.7b00307

Ding, Y., Tang, J., and Guo, F. (2017b). Identification of drug-target interactions via multiple information integration. *Inf. Sci.* 418, 546–560. doi: 10.1016/j.ins.2017.08.045

Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8700–8704. doi: 10.1073/pnas.92.19.8700

Feng, C. Q., Zhang, Z. Y., Zhu, X. J., Lin, Y., Chen, W., Tang, H., et al. (2018). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics.* 2018:bty827. doi: 10.1093/bioinformatics/bty827

Feng, P. M., Chen, W., Lin, H., and Chou, K. C. (2013a). iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* 442, 118–125. doi: 10.1016/j.ab.2013.05.024

Feng, P. M., Ding, H., Chen, W., and Lin, H. (2013b). Naïve Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* 2013:530696. doi: 10.1155/2013/530696

Feng, P. M., Lin, H., and Chen, W. (2013c). Identification of antioxidants from sequence information using naïve Bayes. *Comput. Math. Methods Med.* 2013, 1–5. doi: 10.1155/2013/567529

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

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. (2006). "A closer look at skip-gram modelling," in *Proceedings of the 5th International Conference on Language Resources and Evaluation* (Genoa: LREC), 1–4.

Hang, X., Zeng, W., Zeng, X., and Yen, G. G. (2018). An evolutionary algorithm based on minkowski distance for many-objective optimization. *IEEE Trans. Cybern.* 99, 1–12. doi: 10.1109/TCYB.2018.2856208

Haq, I. U., Chaudhry, W. N., Akhtar, M. N., Andleeb, S., and Qadri, I. (2012). Bacteriophages and their implications on future biotechnology: a review. *Virol. J.*, 9:9. doi: 10.1186/1743-422X-9-9

Hershey, A. D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36, 39–56. doi: 10.1085/jgp.36.1.39

Hu, Y., Zhao, T., Zhang, N., Zang, T., Zhang, J., and Cheng, L. (2018). Identifying diseases-related metabolites using random walk. *BMC Bioinform.* 19(Suppl. 5):116. doi: 10.1186/s12859-018-2098-1

Huang, L., Li, X., Guo, P., Yao, Y., Liao, B., Zhang, W., et al. (2017). Matrix completion with side information and its applications in predicting the antigenicity of influenza viruses. *Bioinformatics* 33, 3195–3201. doi: 10.1093/bioinformatics/btx390

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K. C. (2016). iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. *Molecules* 21:95. doi: 10.3390/molecules21010095

Jiang, S., Liu, B., and Zou, Q. (2018). HITS-PR-HHblits: protein remote homology detection by combining PageRank and hyperlink-induced topic search. *Brief. Bioinform.* 2018:bby104. doi: 10.1093/bib/bby104

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., et al. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.* 17:184. doi: 10.1186/s13059-016-1037-6

Jingjing, H., Ting, F., Zizheng, Z., Bei, H., Xiaolei, Z., and Yi, X. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinform.* 19:306. doi: 10.1186/s12859-018-2321-0

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices11Edited by G. Von Heijne. *J. Mol. Biol.* 292, 195–202. doi: 10.1006/jmbi.1999.3091

Larson, R. E., and Edwards, B. H. (1991). *Elementary Linear Algebra.* 2nd Edn. Lexington, MA: D.C. Heath and Company.

Leyi, W., Huangrong, C., and Ran, S. (2018). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther.* 2018, 635–644. doi: 10.1016/j.omtn.2018.07.004

Leyi, W., Minghong, L., Xing, G., and Quan, Z. (2015). An improved protein structural classes prediction method by incorporating both sequence and structure information. *IEEE Trans. Nanobiosci.* 14, 339–349. doi: 10.1109/TNB.2014.2352454

Li, J., Halgamuge, S. K., Kells, C. I., and Tang, S. L. (2007). Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages. *BMC Bioinform.* 8:S6. doi: 10.1186/1471-2105-8-S4-S6

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., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17, 282–283. doi: 10.1093/bioinformatics/17.3.282

Li, Z., Tang, J., and Guo, F. (2016). Learning from real imbalanced data of 14-3-3 proteins binding specificity. *Neurocomputing* 217, 83–91. doi: 10.1016/j.neucom.2016.03.093

Liu, B. (2017). BioSeq-analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 2017:bbx165. doi: 10.1093/bib/bbx165

Liu, Y., Wang, X., and Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* 20, 330–346. doi: 10.1093/bib/bbx126

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2016). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 905–915. doi: 10.1109/TCBB.2016.2550432

Marks, T., and Sharp, R. (2015). Bacteriophages and biotechnology: a review. *J. Chem. Technol. Biotechnol.* 75, 6–17. doi: 10.1002/(SICI)1097-4660(200001)75:1<6::AID-JCTB157>3.0.CO;2-A

Marvin, D. A., Hale, R. D., Nave, C., and Helmer-Citterich, M. (1994). Molecular models and structural comparisons of native and mutant class I filamentous bacteriophages Ff (fd, f1, M13), If1 and IKe. *J. Mol. Biol.* 235, 260–286. doi: 10.1016/S0022-2836(05)80032-4

Mrozek, D., Daniłowicz, P., and Małysiak-Mrozek, B. (2016). HDInsight4PSi: Boosting performance of 3D protein structure similarity searching with HDInsight clusters in Microsoft Azure cloud. *Inf. Sci.* 349, 77–101. doi: 10.1016/j.ins.2016.02.029

Mrozek, D., Socha, B., Kozielski, S., and Małysiak-Mrozek, B. (2015). An efficient and flexible scanning of databases of protein secondary structures. *J. Intell. Inf. Syst.* 46, 213–233. doi: 10.1007/s10844-014-0353-0

Pearson, K. (1909). Determination of the coefficient of correlation. *Science* 30, 23–25. doi: 10.1126/science.30.757.23

Qiao, Y., Xiong, Y., Gao, H., Zhu, X., and Chen, P. (2018). Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. *BMC Bioinform.* 19:14. doi: 10.1186/s12859-018-2009-5

Qu, K., Han, K., Wu, S., Wang, G., and Wei, L. (2017). Identification of DNA-binding proteins using mixed feature representation methods. *Molecules* 22:E1602. doi: 10.3390/molecules22101602

Robert, C. (2012). Machine learning, a probabilistic perspective. *Chance* 27, 62–63. doi: 10.1080/09332480.2012.726570

Rogers, D. J., and Tanimoto, T. T. (1960). A computer program for classifying plants. *Science* 132, 1115–1118. doi: 10.1126/science.132.3434.1115

Rolf, A. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 32, D115–D119. doi: 10.1093/nar/gkh131

Seguritan, V., Alves, N. Jr., Arnoult, M., Raymond, A., Lorimer, D., Burgin, A. B. Jr., et al. (2012). Artificial neural networks trained to detect

viral and phage structural proteins. *PLoS Comput. Biol.* 8:e1002657. doi: 10.1371/journal.pcbi.1002657

Shen, H. B., and Chou, K. C. (2008). PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* 373, 386–388. doi: 10.1016/j.ab.2007.10.012

Shen, Y., Tang, J., and Guo, F. (2019). Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. *J. Theor. Biol.* 462, 230–239. doi: 10.1016/j.jtbi.2018.11.012

Song, T., Rodríguez-Patón, A., Zheng, P., and Zeng, X. (2018a). Spiking neural P systems with colored spikes. *IEEE Trans. Cogn. Dev. Syst.* 10, 1106–1115. doi: 10.1109/TCDS.2017.2785332

Song, T., Zeng, X., Zheng, P., Jiang, M., and Rodríguez-Patón, A. (2018b). A parallel workflow pattern modeling using spiking neural p systems with colored spikes. *IEEE Trans. Nanobiosci.* 17, 474–484. doi: 10.1109/TNB.2018.2873221

Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., et al. (2018). Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.* doi: 10.2174/1389200219666180820112457. [Epub ahead of print].

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2018). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data*IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2018.2858756. [Epub ahead of print].

Tan, P. N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*, Boston, MA: Pearson Addison Wesley.

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Ting, H., Guangyong, Z., Pingyu, Z., Jia, J., Jing, L., Lu, X., et al. (2014). LAceP: lysine acetylation site prediction using logistic regression classifiers. *PLoS ONE* 9:e89575. doi: 10.1371/journal.pone.0089575

Wang, P., Zhu, W., Liao, B., Cai, L., Peng, L., and Yang, J. (2018). Predicting influenza antigenicity by matrix completion with antigen and antiserum similarity. *Front. Microbiol.* 9:2500. doi: 10.3389/fmicb.2018.02500

Wei, L., Liao, M., Gao, X., and Zou, Q. (2015). Enhanced protein fold prediction method through a novel feature extraction technique. *IEEE Trans. Nanobiosci.* 14, 649–659. doi: 10.1109/TNB.2015.2450233

Wei, L., Su, R., Wang, B., Li, X., Zou, Q., and Gao, X. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites. *Neurocomputing* 324, 3–9. doi: 10.1016/j.neucom.2018.04.082

Wei, L., Tang, J., and Zou, Q. (2017a). SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genomics* 18:742. doi: 10.1186/s12864-017-4128-1

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017b). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med*, 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017c). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wu, L. C., Lee, J. X., Huang, H. D., Liu, B. J., and Horng, J. T. (2009). An expert system to predict protein thermostability using decision tree. *Expert Syst. Appl.* 36, 9007–9014. doi: 10.1016/j.eswa.2008.12.020

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D.-Q. (2018). PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9:2571. doi: 10.3389/fmicb.2018.02571

Xu, L., Liang, G., Shi, S., and Liao, C. (2018a). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Int. J. Mol. Sci.* 19:1773.

Xu, L., Liang, G., Wang, L., and Liao, C. (2018b). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* 9:E158. doi: 10.3390/genes9030158

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., et al. (2017). PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.* 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2'-O-Methylation sites in homo sapiens. *J. Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004

Yang, J., Grunewald, S., and Wan, X. F. (2013). Quartet-net: a quartet-based method to reconstruct phylogenetic networks. *Mol. Biol. Evol.* 30, 1206–1217. doi: 10.1093/molbev/mst040

Yang, J., Grünewald, S., Xu, Y., and Wan, X.-F. (2014). Quartet-based methods to reconstruct phylogenetic networks. *BMC Syst. Biol.* 8, 21–21. doi: 10.1186/1752-0509-8-21

Yang, R., Zhang, C., Gao, R., and Zhang, L. (2015). An ensemble method with hybrid features to identify extracellular matrix proteins. *PLoS ONE* 10:e0117804. doi: 10.1371/journal.pone.0117804

Yao, Y., Li, X., Liao, B., Huang, L., He, P., Wang, F., et al. (2017). Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method. *Sci. Rep.* 7:1545. doi: 10.1038/s41598-017-01699-z

Yi, X., Juan, L., and Dong-Qing, W. (2011). An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins Struct. Funct. Bioinform.* 79, 509–517. doi: 10.1002/prot.22898

Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., and Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomics* 8:S2. doi: 10.1186/1755-8794-8-S2-S2

Yu, L., Ma, X., Zhang, L., Zhang, J., and Gao, L. (2016a). Prediction of new drug indications based on clinical data and network modularity. *Sci. Rep.* 6:32530. doi: 10.1038/srep32530

Yu, L., Su, R., Wang, B., Zhang, L., Zou, Y., Zhang, J., et al. (2017a). Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 966–977. doi: 10.1109/TCBB.2016.2550453

Yu, L., Wang, B., Ma, X., and Gao, L. (2016b). The extraction of drug-disease correlations based on module distance in incomplete human interactome. *BMC Syst. Biol.* 10:111. doi: 10.1186/s12918-016-0364-2

Yu, L., Zhao, J., and Gao, L. (2017b). Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome. *Artif. Intell. Med.* 77, 53–63. doi: 10.1016/j.artmed.2017.03.009

Yu, L., Zhao, J., and Gao, L. (2018). Predicting Potential Drugs for Breast Cancer based on miRNA and Tissue Specificity. *Int. J. Biol. Sci.* 14, 971–982. doi: 10.7150/ijbs.23350

Zeng, X., Ding, N., Rodríguez-Patón, A., and Zou, Q. (2017). Probability-based collaborative filtering model for predicting gene disease associations. *BMC Med. Genomics* 10:76. doi: 10.1186/s12920-017-0313-y

Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 34, 2425–2432. doi: 10.1093/bioinformatics/bty112

Zhang, J., Ju, Y., Lu, H., Xuan, P., and Zou, Q. (2016). Accurate Identification of cancerlectins through hybrid machine learning technology. *Int. J. Genomics* 2016, 1–11. doi: 10.1155/2016/7604641

Zhang, J., and Liu, B. (2017). PSFM-DBT: identifying DNA-binding proteins by combing position specific frequency matrix and distance-bigram transformation. *Int. J. Mol. Sci.* 18:E1856. doi: 10.3390/ijms18091856

Zhang, L., Zhang, C., Gao, R., and Yang, R. (2015). An ensemble method to distinguish bacteriophage virion from non-virion proteins based on protein sequence characteristics. *Int. J. Mol. Sci.* 16, 21734–21758. doi: 10.3390/ijms160921734

Zhu, P., Hu, Q., Han, Y., Zhang, C., and Du, Y. (2016). Combining neighborhood separable subspaces for classification via sparsity regularized optimization. *Inf. Sci.* 370, 270–287. doi: 10.1016/j.ins.2016.08.004

Zhu, P., Hu, Q., Hu, Q., Zhang, C., and Feng, Z. (2018c). Multi-view label embedding. *Pattern Recognit.* 84, 126–135. doi: 10.1016/j.patcog.2018.07.009

Zhu, P., Xu, Q., Hu, Q., and Zhang, C. (2018a). Co-regularized unsupervised feature selection. *Neurocomputing* 275, 2855–2863. doi: 10.1016/j.neucom.2017.11.061

Zhu, P., Xu, Q., Hu, Q., Zhang, C., and Zhao, H. (2018b). Multi-label feature selection with missing labels. *Pattern Recognit.* 74, 488–502. doi: 10.1016/j.patcog.2017.09.036

Zhu, X.-J., Feng, C.-Q., Lai, H.-Y., Chen, W., and Hao, L. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowledge Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Li, X. B., Jiang, W. R., Lin, Z. Y., Li, G. L., and Chen, K. (2014). Survey of MapReduce frame operation in bioinformatics. *Brief. Bioinform.* 15, 637–647. doi: 10.1093/bib/bbs088

Zou, Q., Wang, Z., Guan, X., Liu, B., Wu, Y., and Lin, Z. (2013). An approach for identifying cytokines based on a novel ensemble classifier. *Biomed Res. Int.* 2013:686090. doi: 10.1155/2013/686090

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Conflict of Interest Statement: 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.

Copyright © 2019 Ru, Li and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 09 April 2019
doi: 10.3389/fmicb.2019.00676

A Novel Human Microbe-Disease Association Prediction Method Based on the Bidirectional Weighted Network

Hao Li[1], *Yuqi Wang*[1], *Jingwu Jiang*[2], *Haochen Zhao*[1], *Xiang Feng*[3], *Bihai Zhao*[3] and *Lei Wang*[1,3*]

[1] Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan, China, [2] Clinical Lab, Yongcheng People's Hospital, Shangqiu, China, [3] College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, China

The survival of human beings is inseparable from microbes. More and more studies have proved that microbes can affect human physiological processes in various aspects and are closely related to some human diseases. In this paper, based on known microbe-disease associations, a bidirectional weighted network was constructed by integrating the schemes of normalized Gaussian interactions and bidirectional recommendations firstly. And then, based on the newly constructed bidirectional network, a computational model called BWNMHMDA was developed to predict potential relationships between microbes and diseases. Finally, in order to evaluate the superiority of the new prediction model BWNMHMDA, the framework of LOOCV and 5-fold cross validation were implemented, and simulation results indicated that BWNMHMDA could achieve reliable AUCs of 0.9127 and 0.8967 ± 0.0027 in these two different frameworks respectively, which is outperformed some state-of-the-art methods. Moreover, case studies of asthma, colorectal carcinoma, and chronic obstructive pulmonary disease were implemented to further estimate the performance of BWNMHMDA. Experimental results showed that there are 10, 9, and 8 out of the top 10 predicted microbes having been confirmed by related literature in these three kinds of case studies separately, which also demonstrated that our new model BWNMHMDA could achieve satisfying prediction performance.

Keywords: microbe, disease, association prediction, bidirectional weighted network, bidirectional recommendations

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Yan Zhao,
China University of Mining and Technology, China
Jincai Yang,
Central China Normal University, China
Xinguo Lu,
Hunan University, China

***Correspondence:**
Lei Wang
wanglei@xtu.edu.cn

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

Received: *19 December 2018*
Accepted: *18 March 2019*
Published: *09 April 2019*

Citation:
Li H, Wang Y, Jiang J, Zhao H, Feng X, Zhao B and Wang L (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

1. INTRODUCTION

Microorganisms are small in shape, simple in structure, and closely related to human beings. The development of modern bioinformatics and sequencing technologies has led to the study of microorganisms living in the ocean, soil, human body, and other places by the scientific community (Gilbert and Dupont, 2011). Among them, eukaryotes, archea, bacteria, and viruses are human-related microorganisms, collectively known as human microbiota (Turnbaugh et al., 2007; Methé et al., 2012). Microorganisms exist in large quantities in humans, nearly 10 times that of human cells (Sender et al., 2016). According to recent researches, there are nearly 1,014 bacterial

cells in the human body with more than 10,000 kinds of microorganisms, which provide different degrees of metabolic activity (Bhavsar et al., 2007; Turnbaugh et al., 2007; Shah et al., 2016). Parasitic in the human body, these microbes do not harm the host, but are interdependent with human beings and are called "forgotten organs" (Quigley, 2013). With the continuous advancement of high-throughput sequencing technology and analytical systems, people have gradually realized the importance of microorganisms in the investigation. According to the survey, microbes participate in a series of human life activities, such as harvesting and storing energy, regulating the immune system, protecting the human body from foreign microorganisms and pathogens, participating in the digestion and absorption of carbohydrates and promoting metabolism (Guarner and Malagelada, 2003; Gill et al., 2006). Therefore, once the microbes become "unhealthy" in the human body, the human body will receive their effects leading to physiological disorders and even illness.

Humans and commensal microbiota have formed a close symbiotic relationship in the process of continuous evolution. The microbiota will be affected by the host and living environment. It has been reported that diet affects the structure and activity of human intestinal microbes (Duncan et al., 2006; Ley et al., 2006; Walker et al., 2010; David et al., 2013) For example, a short-term high-fat, low-fiber diet can cause changes in microbial structure, while long-term diets are associated with alternative intestinal status (Wu et al., 2011). Besides, smoking (Mason et al., 2014), age, and genes are also factors influencing the composition of the microbiota (Gill et al., 2006). Therefore, once the human body and the microbiota cannot coexist harmoniously, it may cause various problems in the human body. Based on the 16S ribosomal RNA (rRNA) gene sequence and classification spectrum (Thompson et al., 2014; Jesmok et al., 2016), researchers have found that a large number of human diseases are closely related to human microorganisms, including cancer (Moore and Moore, 1995), diabetes (Wen et al., 2008; Brown et al., 2011; Qin et al., 2012), Obesity (Ley et al., 2005; Zhang et al., 2009), kidney stones (Hoppe et al., 2011), and other thorny diseases. For example, Huang (2013) pointed out that microbes can affect allergic sensitization and asthma development in susceptible individuals, and early intervention in promoting "healthy" human microbiome constitution may have the potential and benefits of preventing asthma. Hence, some researchers are proposing to promote the induction of sensitized immune response through the research and development of probiotic-based therapies (Rauch and Lynch, 2012).

Disease-related microbes are obtaining more and more attention from humans, and researchers have carried out some large-scale sequencing projects, including the Human Microbiome Project (HMP) (Turnbaugh et al., 2007) and the Earth Microbiome Project (EMP) (Gilbert et al., 2010). Moreover, some databases (Matsumoto et al., 2005; Faith et al., 2007; Chen et al., 2010; Mikaelyan et al., 2015) for categorizing and managing disease-related microbial information have also been developed. For instance, Ma et al. collected and compiled 483 pairs of human microbe-disease associations by collecting published literature and established the Human Microbe-Disease Association Database (HMDAD) (Ma et al., 2016). These accurate data provide the possibility to predict human microbes and diseases. Nowadays, most microbial community identification methods are independent culture methods and quantitative methods. Their shortcomings are obvious and often take a lot of time and efforts. Previously, many researchers have studied the potential correlation predictions of diseases and other biological categories (such as miRNA Chen and Yan, 2014; You et al., 2017; Chen et al., 2018b,c and lncRNA Chen and Yan, 2013; Chen et al., 2016b, 2018a; Yu et al., 2018; Xuan et al., 2019), and simultaneously, Drug-target interaction prediction (Chen et al., 2012) and the study of synergistic drug combinations prediction (Chen et al., 2016a) has also achieved satisfying successes. And among existing state-of-the-art methods, the computational model of KATZ measure for human microbe-disease association prediction (KATZHMDA) (Chen et al., 2017) proposed by Chen et al. is one of their prominent representatives, which not only achieved excellent prediction performance but also initialized the research field of the microbe-disease prediction. Later, Huang Z.A. et al. (2017) proposed a Path-Based computational model of Human Microbe-Disease Association prediction (PBHMDA), which adopts a special depth-first search algorithm to traverse all possible paths between microbes and diseases in heterogeneous networks to obtain the prediction score of each microbe-disease pair. Wang et al. (2017) proposed a semi-supervised learning-based computational model of Laplacian Regularized Least Squares for Human Microbe-Disease Association prediction (LRLSHMDA), which utilizes Laplace's regular least squares classification combined with topological information of the known microbe-disease association network to train an optimal classifier. Huang Y.A. et al. (2017) developed a method based on Neighbor and Graph-based combined recommendation model for Human Microbe-Disease Association prediction (NGRHMDA) by combining two recommendation models as a neighbor-based collaborative filtering model and a topology-based model. Peng et al. (2018) developed a model of Adaptive Boosting for Human Microbe-Disease Association prediction (ABHMDA), which reveals the associations between disease and microbe by using a strong classifier to calculate the probability of disease-microbe pair association. In addition, Shen et al. (2018) proposed Bi-Random Walk based on Multiple Path (BiRWMP) to predict microbe-disease associations. Shi et al. (2018) propose BMCMDA based on Binary Matrix Completion to predict potential microbe-disease associations.

In this paper, inspired by the performance of KATZHMDA, we proposed a new microbe-disease association prediction model called BWNMHMDA. A novel two-way network was constructed firstly based on the known microbe-disease associations downloaded from the HMDAD database, and then, the Gaussian interaction profile kernel similarity were adopted to assign weights to every node and edge in a newly constructed two-way network. Hence, a bidirectional weighted network was further obtained by implementing two newly developed bidirectional recommendation measures. Finally, based on the newly constructed bidirectional weighted network, a computational model was constructed to infer potential

microbe-disease associations. In order to estimate the prediction performances of BWNMHMDA, the framework of leave-one-out cross validation (LOOCV) and 5-fold cross validation(5-Fold CV) were implemented, and simulation results indicated that BWNMHMDA could achieve reliable AUCs of 0.9127 in LOOCV and 0.8967 ± 0.0027 in 5-Fold CV, respectively, which is much better than that of state-of-the-art methods. And moreover, in case studies of asthma, colorectal carcinoma, and chronic obstructive pulmonary disease, the simulation results also demonstrated the effective predictability of BWNMHMDA.

2. MATERIAL

Since known microbe-disease associations were considered in our prediction model BWNMHMDA, we firstly downloaded known microbe-disease associations from the Human Microbe-Disease Association database (HMDAD) (Ma et al., 2016), and as a result, after getting rid of the redundant associations, a total of 450 different microbe-disease associations including 39 human diseases and 292 microbes were collected from 61 public publications. Hence, a 39×292 dimensional adjacency matrix A is obtained finally, which will be utilized as the data source of our prediction model BWNMHMDA. And additionally, in the adjacency matrix A, the value of $A[i][j]$ is set to 1 if there is a known association between the ith disease and the jth microbe, otherwise, $A[i][j]$ is set to 0.

3. METHODS

As illustrated in the following **Figure 1**, in BWNMHMDA, three kinds of association networks such as the known microbe-disease association network, the microbe similarity network and the diseases similarity network will be constructed firstly. And then, through integrating these three kinds of association networks, an integrated microbe-disease heterogeneous association network will be obtained. Moreover, through adopting the Gaussian interaction profile kernel similarity to assign weights to every node and edge in the integrated microbe-disease heterogeneous association network, a bidirectional weighted microbe-disease association network can be further obtained. Hence, based on the newly constructed bidirectional weighted association network, a novel computation model can be developed to infer potential microbe-disease associations.

3.1. Microbes Similarity Based on Gaussian Interaction Profile Kernel Similarity

It is obviously reasonable that for any two microbes if there are more common human diseases proved to be related to them, may tend to share more functional similarities potentially. Hence, in the known microbe-disease association network, we will first adopt the Gaussian interaction profile kernel similarity to construct a microbe similarity network according to the following formula (1):

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

Where $m(i)$ and $m(j)$ represent the ith and jth microbes respectively in the adjacency matrix A, $IP[m(i)]$ and $IP[m(j)]$ denote ith and jth column, respectively, in the adjacency matrix A, and $\|X\|$ represents the norm of the vector X. Moreover, the parameter γ_m can be obtained as follows:

$$\gamma_m = \gamma_m{}' / \frac{1}{N_m} \sum_{i=1}^{N_m} \left\| IP(m(i)) \right\|^2 \quad (2)$$

Here, $\gamma_m{}'$ is a parameter utilized to control the Gaussian kernel bandwidth, and according to the related studies (van Laarhoven et al., 2011), $\gamma_m{}'$ will be set to 1 in BWNMHMDA. In addition, the parameter N_m indicates the total number of microbes collected from the HMDAD database, and it is obvious that there is N_m=292.

Thereafter, according to the above formula (1), it is easy to see that a microbe similarity matrix KM can be calculated, specifically, and for simplicity, we will replace $KM[m(i), m(j)]$ with $KM(i, j)$ in the following sections.

3.2. Diseases Similarity Based on Gaussian Interaction Profile Kernel Similarity

In a similar way, through adopting the Gaussian interaction profile kernel similarity, we can further construct a disease similarity network according to the following formula (3):

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

Here, the parameter γ_d can be obtained as follows:

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

Here, $\gamma_d{}'$ is a parameter utilized to control the Gaussian kernel bandwidth, and according to the related studies (van Laarhoven et al., 2011), $\gamma_d{}'$ will be also set to 1. In addition, the parameter N_d indicates the total number of diseases collected from the HMDAD database, and it is obvious that there is N_d=39.

Thereafter, according to the above formula (3), it is easy to see that a disease similarity matrix KD can be calculated, specifically, and for simplicity, we will replace $KD[d(i), d(j)]$ with $KD(i, j)$ in the following sections.

3.3. Data Pre-processing

Based on the newly constructed microbe similarity network and disease similarity network, after integrating the known microbe-disease associations with these two similarity networks, it is obvious that we can construct an integrated heterogeneous microbe-disease association network consisting of two kinds of nodes such as microbe and disease, and three kinds of edges such as the edges between microbes, the edges between microbes and diseases, and the edges between diseases. And furthermore, based on the integrated heterogeneous microbe-disease association network, we can obtain a (39+292)×(39+292) dimensional matrix P as follows:

$$p = \begin{bmatrix} KD & A \\ A^T & KM \end{bmatrix} \quad (5)$$

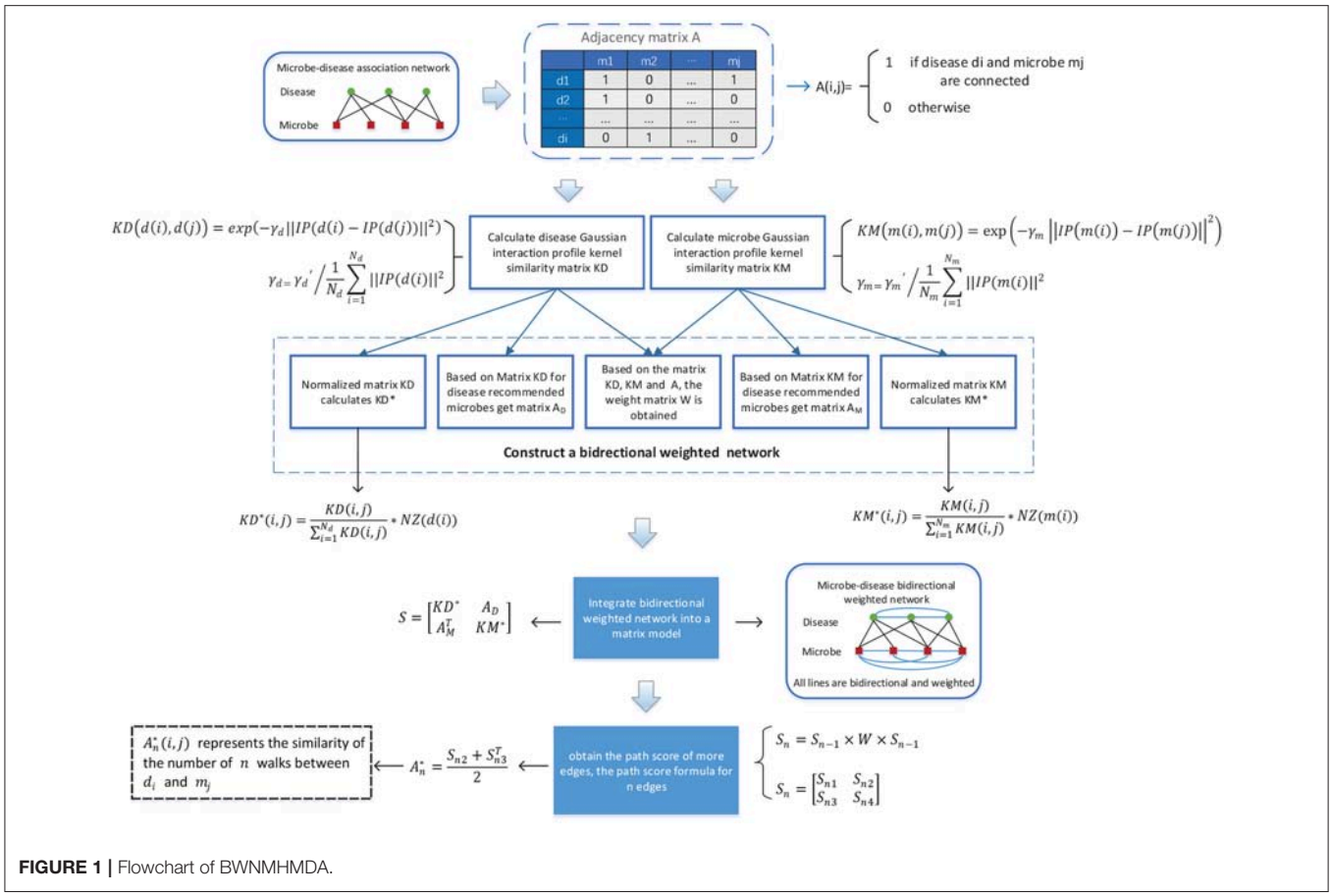


FIGURE 1 | Flowchart of BWNMHMDA.

Moreover, in the integrated heterogeneous microbe-disease association network, if a microbe (or disease) node has more edges connecting with disease (or microbe) nodes, then it is obvious that the microbe (or disease) node will have less significance to those disease (or microbe) nodes connecting with it, which means that the microbe (or disease) node shall be assigned smaller weights than those microbe (or disease) nodes with fewer edges. Hence, based on above formula (5), we can further obtain a (39+292)×(39+292) dimensional diagonal matrix W to represent the weight value of each node in the heterogeneous network as follows:

$$W = diag(1/(P \times P^T)) \tag{6}$$

In addition, while calculating the similarity between two nodes in the heterogeneous network, there may be cases where the scores of the path consisting of three edges are larger than the scores of the path consisting of two edges. Hence, in order to avoid such kind of situation, we will normalize the weights of edges in the heterogeneous network by adopting the following formula (7) and formula (8) separately.

$$KM^*(i,j) = \frac{KM(i,j)}{\sum_{i=1}^{N_m} KM(i,j)} \times NZ(m(i)) \tag{7}$$

Where $NZ[m(i)]$ denotes the number of elements with non-zero values in the ith row of the matrix KM. And based on above formula (7), it is noteworthy that the symmetric matrix KM will be changed to an asymmetric matrix KM^* after the normalization. Moreover, in the heterogeneous network, $KM^*(i,j)$ represents the weight of the directed edge from the microbe node m_i to the microbe node m_j, while $KM^*(j,i)$ denotes the weight of the directed edge from the microbe node m_j to the microbe node m_i.

$$KD^*(i,j) = \frac{KD(i,j)}{\sum_{i=1}^{N_d} KD(i,j)} \times NZ(d(i)) \tag{8}$$

Where $NZ[d(i)]$ denotes the number of elements with non-zero values in the ith row of the matrix KD. And based on the above formula (8), it is noteworthy that the symmetric matrix KD will as well be changed to an asymmetric matrix KD^* after the normalization. Moreover, in the heterogeneous network, $KD^*(i,j)$ represents the weight of the directed edge from the disease node d_i to the disease node d_j, while $KD^*(j,i)$ denotes the weight of the directed edge from the disease node d_j to the disease node d_i .

Therefore, according to the above descriptions, it is obvious that we can obtain a bidirectional heterogeneous network based on the above formula (7) and formula (8).

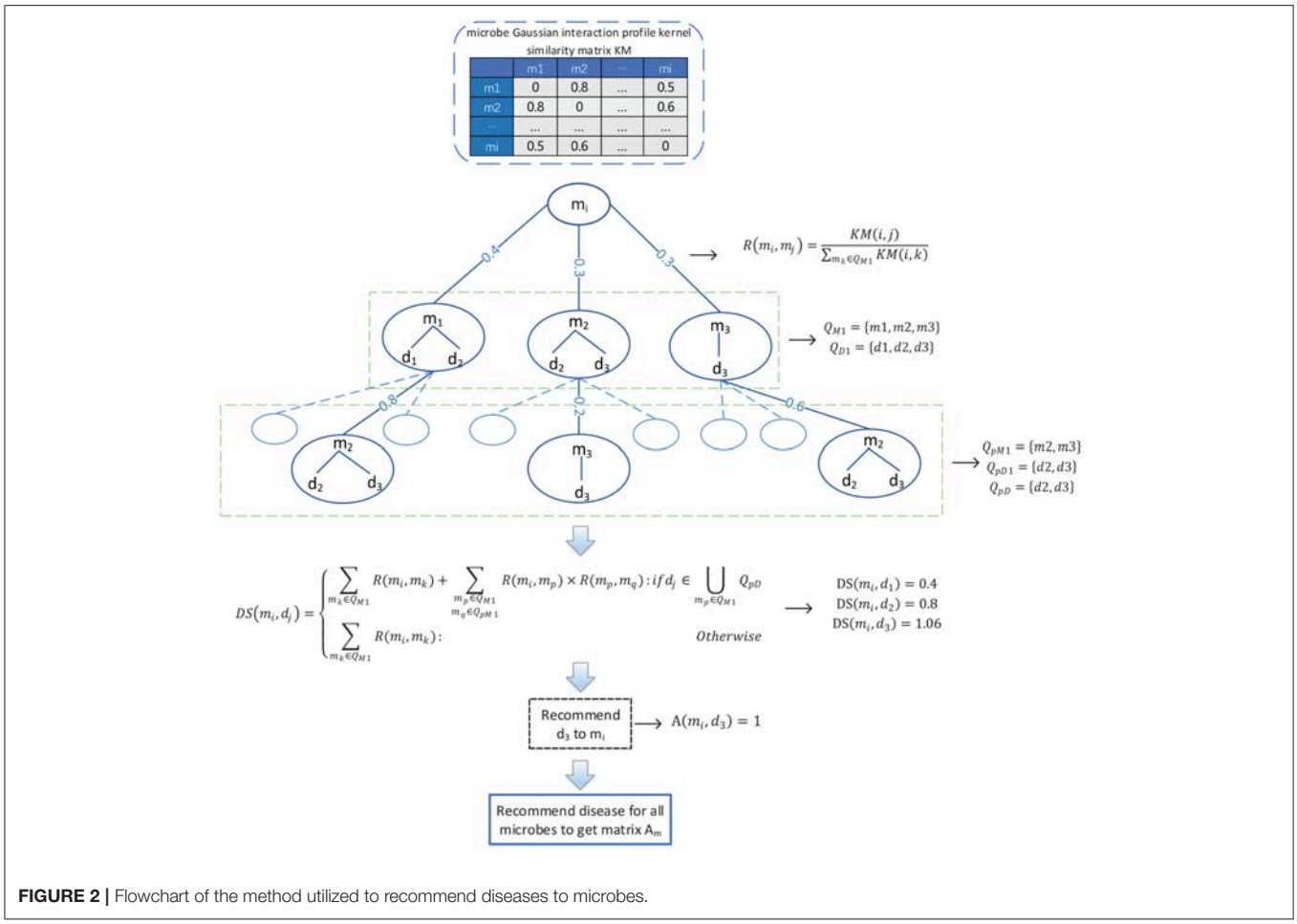


FIGURE 2 | Flowchart of the method utilized to recommend diseases to microbes.

3.4. Bidirectional Recommendation of Potential Associations

Considering that there are only 450 known associations in the adjacency matrix A, which is very sparse, therefore, in order to solve the problem of the adjacency matrix A caused by the scarcity of known associations, as illustrated in the following **Figure 2**, we designed a novel bidirectional recommendation model in this section based on the bidirectional heterogeneous network constructed above. And in this bidirectional recommendation model, we first designed a recommendation algorithm to recommend diseases for microbes based on the Gaussian interaction profile kernel similarities between microbes as follows:

(1) Firstly, for any given microbe node m_i in the bidirectional heterogeneous network, let Q_{M1} denote the set consisting of the first K microbes that are other than m_i in the bidirectional heterogeneous network and most similar to m_i at the same time, and considering about the time complexity, in this paper, K will be set to 3. And then, let Q_{D1} represent the set of diseases having known associations with at least one of the microbe nodes in Q_{M1}, thereafter for any microbe node m_j in Q_{M1}, we can obtain the recommendation score of m_j to m_i according to the following formula (9):

$$R(m_i, m_j) = \frac{KM(i,j)}{\sum_{m_k \in Q_{M1}} KM(i,k)} \tag{9}$$

Moreover, for any given disease node d_j in Q_{D1}, we can further obtain the recommendation score of d_j to m_i according to the following formula (10):

$$DS(m_i, d_j) = \sum_{m_k \in Q_{M1}} R(m_i, m_k) \tag{10}$$

Hence, in a similar way, for any given microbe node m_p in Q_{M1}, we can obtain a set Q_{pM1} consisting of the first K microbes that are other than m_p in the bidirectional heterogeneous network and most similar to mp at the same time, and then, based on the set Q_{pM1}, we can further obtain a set Q_{pD1} consisting of diseases that have known associations with at least one of the microbe nodes in Q_{pM1}. In addition, let $Q_{pD} = Q_{D1} \cap Q_{pD1}$, it is obvious that for any node d_k in $\cup_{m_p \in Q_{M1}} Q_{pD}$, it shall be assigned higher recommendation score than those nodes that are in Q_{D1} and not in $\cup_{m_p \in Q_{M1}} Q_{pD}$. Hence, for any given disease node d_j in Q_{D1}, based on the above formula

(10), we can obtain a modified recommendation score of d_j to m_i as follows:

$$DS(m_i, d_j) = \begin{cases} \sum\limits_{m_k \in Q_{M1}} R(m_i, m_k) + \sum\limits_{\substack{m_p \in Q_{M1} \\ m_q \in Q_{pM1}}} R(m_i, m_p) \times R(m_p, m_q) : if\ d_j \in \bigcup\limits_{m_p \in Q_{M1}} Q_{pD} \\ \sum\limits_{m_k \in Q_{M1}} R(m_i, m_k) : otherwise \end{cases} \tag{11}$$

Obviously, according to the above formula (11), for all these disease nodes in Q_{D1}, we can obtain their corresponding recommendation scores, after sorting these disease nodes according to their recommendation scores in descending order, we will finally recommend the disease node ranking first to the microbe node m_i. And additionally, for the microbe node m_i, supposing that the disease node that we recommended to it is d_j, then we will further set the value of $A(i, j)$ in the adjacency matrix A to 1. Consequently, through updating the adjacency matrix A as stated above, it is obvious that we can obtain a new adjacency matrix A_m.

(2) Secondly, in a similar way, for any given disease node d_i in the bidirectional heterogeneous network, let Q_{D2} denote the set consisting of the first K (=3) diseases that are other than d_i in the bidirectional heterogeneous network and most similar to d_i at the same time, and then, let Q_{M2} represent the set of microbes having known associations with at least one of the disease nodes in Q_{D2}, thereafter, for any given disease node d_p in Q_{D2}, we can obtain a set Q_{pD2} consisting of the first K diseases that are other than d_p in the bidirectional heterogeneous network and most similar to d_p at the same time. Moreover, based on the set Q_{pD2}, we can further obtain a set Q_{pM2} consisting of microbes that have known associations with at least one of the disease nodes in Q_{pD2}. Finally, let $QpM = Q_{M2} \cap Q_{pM2}$, then for any given microbe node m_j in Q_{M2}, we can obtain a recommendation score of m_j to d_i as follows:

$$DS(d_i, m_j) = \begin{cases} \sum\limits_{d_k \in Q_{D2}} R(d_i, d_k) + \sum\limits_{\substack{d_p \in Q_{D1} \\ d_q \in Q_{pD2}}} R(d_i, d_p) \times R(d_p, d_q) : if\ m_j \in \bigcup\limits_{d_p \in Q_{D2}} Q_{pM} \\ \sum\limits_{d_k \in Q_{D2}} R(d_i, d_k) : otherwise \end{cases} \tag{12}$$

Here,

$$R(d_i, d_j) = \frac{KD(i, j)}{\sum_{d_k \in Q_{D2}} KD(i, k)} \tag{13}$$

Obviously, according to the above formula (12), for all these microbe nodes in Q_{M2}, we can obtain their corresponding recommendation scores, after sorting these microbe nodes according to their recommendation scores in descending order, we will finally recommend the microbe node ranking first to the disease node d_i. And additionally, for the disease node d_i, supposing that the microbe node that we recommended to it is m_j, then we will further set the value of $A(j, i)$ in the adjacency matrix A to 1. Consequently, through updating the adjacency matrix A as stated above, it is obvious that we can obtain a new adjacency matrix A_d.

3.5. Prediction Model of BWNMHMDA

KATZ is a network-based method that can solve link prediction problems. In recent years, KATZ has been implemented successfully in many different prediction applications such as prediction of social networks (Katz, 1953), prediction of associations between gene (Yang et al., 2014) and prediction of associations between lncRNAs (Chen, 2015), etc. In 2017, Chen et al. further applied KATZ in the field of microbe-disease association prediction for the first time (Chen et al., 2017). Considering that KATZ can be utilized to calculate the similarities between nodes in heterogeneous networks, and according to the above description in section 3.3, we have built a bidirectional heterogeneous microbe-disease association network, hence, in this section, we will design a model called BWNMHMDA based on KATZ to predict potential microbe-disease associations. For constructing the prediction model, we will convert the bidirectional heterogeneous microbe-disease association network to a (39+292)*(39+292) dimensional matrix S as follows:

$$S = \begin{bmatrix} KD^* & A_d \\ A_m^T & KM^* \end{bmatrix} \tag{14}$$

Hence, based on above formula (14), for any given disease node d_i and microbe node m_j in the bidirectional heterogeneous microbe-disease association network, we can predict the potential similarity between them as follows:

$$Sim(d_i, m_j) = A_n^*(i, j) \tag{15}$$

Here, n is a parameter representing the number of steps between disease nodes and microbe nodes in the bidirectional heterogeneous microbe-disease association network. For n = 1, 2, 3, ..., there are:

$$A_n^* = \frac{S_{n2} + S_{n3}^T}{2} \tag{16}$$

$$S_n = S_{n-1} \times W \times S_{n-1} = \begin{bmatrix} S_{n1} & S_{n2} \\ S_{n3} & S_{n4} \end{bmatrix} \tag{17}$$

$$S_2 = S \times W \times S \tag{18}$$

Specifically, in formula (16), the matrix $S_{n2}(i, j)$ represents the total score of all paths with length of n from the disease d_i to microbe m_j, and correspondingly, the matrix $S_{n3}(j, i)$ represents the total score of all paths with length of n from the microbe m_j to disease d_i. It is worth noting that since the weights of the edges

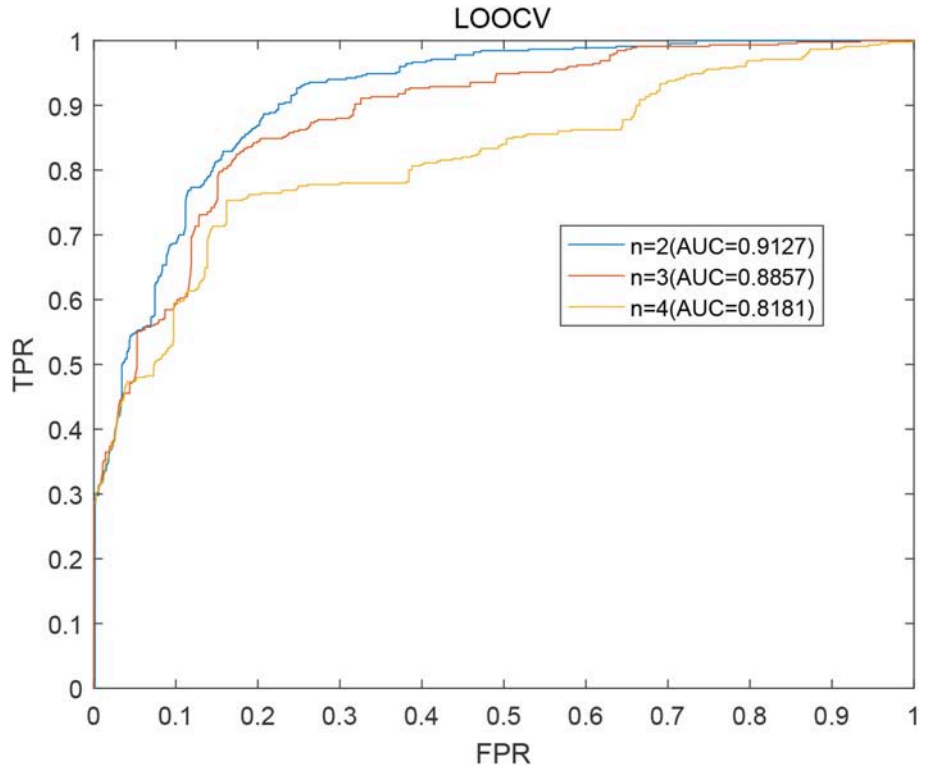


FIGURE 3 | AUCs achieved by BWNMHMDA in LOOCV while n = 2, 3, 4 separately.

in the heterogeneous network are bidirectional, we integrate S_{n2} and S_{n3} as formula (16). The two matrices are assigned the same weight as the final predictive score matrix A_n^*.

4. RESULT

4.1. Effects of the Parameter *n* to BWNMHMDA

The framework of Leave-one-out cross validation (LOOCV) and 5-fold cross validation (5-Fold CV) are two kinds of common methods to evaluate model performance. While implementing LOOCV on our prediction model BWNMHMDA, each known microbe-disease association will be used as a test sample and further predicted by training the other known microbe-disease associations. Moreover, all microbe-disease pairs without known relevant evidence will be considered as candidate samples. The predicted score which obtained a higher rank than the given threshold will be considered as a successful prediction. Obviously, while setting different thresholds, the true positive rate (TPRs, sensitivity) and false positive rate (FPRs, 1-specificity) can be obtained. Here, sensitivity refers to the percentage between the number of test samples with ranks higher than the given threshold and the number of positive samples (known microbe-disease associations). Meanwhile, 1-specificity denotes the percentage of negative microbe-disease associations which obtained ranks lower than the threshold. Finally, the receiver operating characteristic (ROC) curve can be further drawn. The area under the ROC curve(AUC) can be calculated to evaluate its predictive performance, where the AUC value of 1 indicates perfect prediction perfection and the AUC value of 0.5 implies pure random prediction performance (Chen et al., 2017).

As described above, in our prediction model BWNMHMDA, the variable n in the formulas (15) is a critical parameter. Hence, we will first estimate its effect to the prediction performance of BWNMHMDA in this section. And as illustrated in **Figure 3**. BWNMHMDA achieved the best prediction performance while n = 2, and as the value of n sequentially increased from 2 to 4, the AUCs achieved by BWNMHMDA decreased continuously, and through analysis, we found that the reason may be that the number of known microbe-disease associations is minimal in the HMDAD database, which leads that long paths in the bidirectional heterogeneous microbe-disease association network will be meaningless to the prediction performance of BWNMHMDA.

In order to further evaluate the effects of the parameter n to our prediction model, we further implemented 5-fold cross validation on BWNMHMDA, and during simulation, all known microbe-disease associations were randomly divided into five segments with almost the same size, among which, four segments were utilized for model learning, and the remaining segment were used as test samples for model evaluation. Similar to LOOCV, all microbe-disease pairs without relevant evidence would be considered as potential candidates. In order

TABLE 1 | AUCs achieved by BWNMHMDA in the framework of 5-Fold CV while n = 2, 3, 4 separately.

n = 2	n = 3	n = 4
0.8967 ± 0.0027	0.8804 ± 0.0026	0.8109 ± 0.0052

to reduce the experimental bias, we repeated our simulation based on the 5-fold cross validation 100 times, and during each time of simulation, the samples were divided randomly. Finally, as illustrated in the following **Table 1**, it is easy to see that BWNMHMDA could as well achieve the best prediction performance while n=2, and moreover, as the value of n sequentially increased from 2 to 4, the AUCs achieved by BWNMHMDA also decreased continuously. Hence, we will set n to 2 in the subsequent experiments.

4.2. Comparison With Other State-of-the-Art Methods

In order to verify the prediction performance of BWNMHMDA, in this section, we compared it with KATZHMDA (Chen et al., 2017), BiRWMP (Shen et al., 2018), and LRLSHMDA (Wang et al., 2017) based on the dataset of known microbe-disease associations downloaded from the HMDAD database. And as illustrated in the following **Figure 4** and **Table 2**, it is easy to see that in LOOCV, BWNMHMDA can achieve a reliable AUC of 0.9127 that is much better than the AUC achieved by KATZHMDA (0.8382), BiRWMP (0.8637), and LRLSHMDA (0.8909), and in the framework of 5-fold cross validation, BWNMHMDA can achieve a reliable AUC of 0.8967 ± 0.0027 that is much better than the AUC achieved by KATZHMDA (0.8301 ± 0.0033), BiRWMP (0.8522 ± 0.0054), and LRLSHMDA (0.8794 ± 0.0029) as well.

We further compare BWNMHMDA with NGRHMDA (Huang Y.A. et al., 2017), ABHMDA (Peng et al., 2018), and BMCMDA (Shi et al., 2018) in LOOCV based on the same dataset. As shown in **Table 3**, our method achieves the best performance.

5. CASE STUDIES

In order to further measure the prediction performance of BWNMHMDA, in this section, we selected three kinds of important human diseases such as asthma, colorectal carcinoma, and COPD (Chronic Obstructive Pulmonary Disease) to explore the associations between the human microbes and the human respiratory and digestive system diseases. Among them, asthma is a heterogeneous disease process accompanied by recurrent episodes of wheezing, chest tightness, difficulty breathing, and indirect cough (Busse, 2007). In recent years, the prevalence of asthma is rising rapidly. It is reported that about 8% of people have been affected by asthma by 2010, especially in the children's population (Guilbert et al., 2014). Hence, considering that asthma has been demonstrated to be closely associated with microbes as well (Çalşkan et al., 2013; Gilstrap and Kraft, 2013), for example, *Hemophilia*, *Moraxella*, and *Neisseria* spp. in the lungs of asthma patients are proved to be closely related to the increased risk of asthma in the neonatal oropharynx. *Staphylococcus* was found in the respiratory tract of children with asthma (Sullivan et al., 2016), in this section, we selected asthma as one of our case studies to evaluate the performance of BWNMHMDA. And as illustrated in the following **Table 4**, all of these top 10 microorganisms predicted by BWNMHMDA have been verified to be associated with the onset of asthma. For example, *Tropheryma whipplei* (Ranking first in the list of top 10 predicted microbes) has been confirmed to be abundant in airway of patients with eosinophilic asthma (Simpson et al., 2015). *Clostridium difficile* (Ranking second in the list of top 10 predicted microbes) has been confirmed to be associated with asthma after 6–7 years of colonization (van Nimwegen et al., 2011). *Firmicutes* (Ranking third in the list of top 10 predicted microbes) has been confirmed to be increased in severe asthmatics (Zhang et al., 2016). Furthermore, the increased sensitivity to *Staphylococcus aureus* (Ranking fifth in the list of top 10 predicted microbes) has been proved to be a marker of eosinophilic inflammation and severe asthma in asthmatic patients as well (Nagasaki et al., 2017). We published evidence for the top 10 potential asthma-related microbes predicted by BWNMHMDA in the **Table 4**.

In recent years, colorectal carcinoma (CRC) is becoming a major cause of cancer mortality in both China and the United States. In 2016, an estimated 134,000 people had been diagnosed with CRC, and approximately 49,000 had died of CRC (Bibbins-Domingo et al., 2008). By gender, CRC is the second most common cancer in women (about 9.2%) and the third in men (about 10%) (Astin et al., 2011). Since it has been proved that CRC is related to gut microbiota such as the *Fusobacterium*, the *Bacteroides fragilis* and the enteropathogenic *Escherichia coli*, and the dysbiosis of these gut microbiotas will induce colon cancer through a chronic inflammatory mechanism (Mármol et al., 2017). Hence in this section, we selected CRC as one of our case studies to evaluate the performance of BWNMHMDA. And as illustrated in the following **Table 5**, there are 9 out of these top 10 microorganisms predicted by BWNMHMDA have been verified to be associated with the onset of colorectal carcinoma. For instance, related studies have shown that the abundance of *Firmicutes* (Ranking 6th in the list of top 10 predicted microbes) in the lumen of CRC rats will increase, while the abundance of *Bacteroidetes* (Ranking 4th in the list of top 10 predicted microbes) will reduce. And moreover, the abundance of *Proteobacteria* (Ranking second in the list of top 10 predicted microbes) has been confirmed to be higher in CRC rats than in healthy rats. Meanwhile, *Bacteroides* (Ranking 9th in the list of top 10 predicted microbes) has been proved to of a relatively high abundance in CRC rats at the genus level. *Prevotella* (Ranking third in the list of top 10 predicted microbes) has been found to be significantly more abundant in healthy rats than CRC rats (Zhu et al., 2014). Additionally, compared with the healthy control group, Fukugaiti MH et al. detected more *C. difficile* (Ranking 5th in the list of top 10 predicted microbes) in the cancer group, which suggests that these bacteria may play an important role in the colorectal carcinoma (Fukugaiti et al., 2015). We published evidence for the top 10

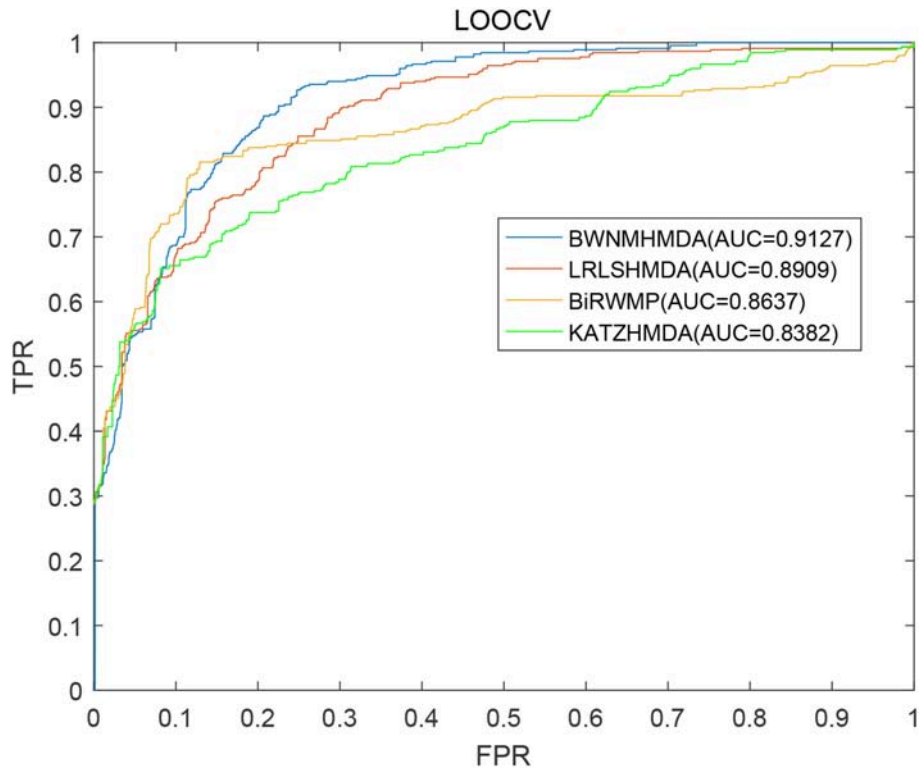


FIGURE 4 | AUCs achieved by KATZHMDA, BiRWMP, LRLSHMDA, and BWNMHMDA in LOOCV.

TABLE 2 | AUCs achieved by BWNMHMDA, KATZHMDA, BiRWMP, and LRLSHMDA in LOOCV and 5-Fold CV separately.

Method	LOOCV	5-Fold CV
BWNMHMDA	0.9127	0.8967 ± 0.0027
LRLSHMDA	0.8909	0.8794 ± 0.0029
BiRWMP	0.8637	0.8522 ± 0.0054
KATZHMDA	0.8382	0.8301 ± 0.0033

TABLE 3 | AUCs achieved by BWNMHMDA, NGRHMDA, ABHMDA, and BMCMDA in LOOCV separately.

Method	BWNMHMDA	NGRHMDA	ABHMDA	BMCMDA
AUC	0.9127	0.8938	0.8869	0.906

TABLE 4 | Top 10 potential asthma-related microbes predicted by BWNMHMDA and all of these 10 microbes have been confirmed by evidences.

Rank	Microbe	Evidence
1	*Tropheryma whipplei*	PMID: 26647445
2	*Clostridium difficile*	PMID: 21872915
3	*Firmicutes*	PMID: 27078029
4	*Lachnospiraceae*	PMID: 26512904
5	*Staphylococcus aureus*	PMID: 17950502
6	*Clostridia*	PMID: 22047069
7	*Bacteroides*	PMID: 18822123
8	*Fusobacterium*	PMID: 24024497
9	*Clostridium coccoides*	PMID: 21477358
10	*Actinobacteria*	PMID: 23265859

potential CRC-related microbes predicted by BWNMHMDA in the **Table 5**.

Finally, COPD is an obstructive pulmonary disease that worsens over time, and the main symptoms of COPD are shortness of breath and coughing. And as of 2015, patients with chronic obstructive pulmonary disease accounted for approximately 174.5 million (about 2.4%) of the global population (Vos et al., 2016). For the past few years, due to high smoking rates and an aging population in developing countries, the death toll of COPD is rising fast (Mathers and Loncar, 2006). Although treatments can slow the progression of COPD, there is no cure yet. Considering that many evidences have demonstrated that there exist associations between microbiomes and COPD, for instance, Galiana et al. found that the microbiota diversity of patients with severe COPD was lower than that of mild/moderate diseases, and actinomyces accounted for a high proportion of patients with severe COPD (Galiana et al., 2013), hence in this section, we selected COPD as one of our case studies to evaluate the performance of BWNMHMDA. And as illustrated in the following **Table 6**, there are 8 out of these top 10 microorganisms predicted by BWNMHMDA have been verified to be associated with the onset of COPD. For instance, COPD has been confirmed to be a kind of essential comorbidity in human immunodeficiency virus (HIV) patients, and more *T.*

TABLE 5 | Top 10 potential CRC-related microbes predicted by BWNMHMDA and 9 out of these 10 microbes have been confirmed by evidences.

Rank	Microbe	Evidence
1	*Tropheryma whipplei*	Unconfirmed
2	*Proteobacteria*	PMID:24603888
3	*Prevotella*	PMID:29432368
4	*Bacteroidetes*	PMID:26992426
5	*Clostridium difficile*	PMID:19807912
6	*Firmicutes*	PMID:29985435
7	*Helicobacter pylori*	PMID:11774957
8	*Clostridia*	PMID:26691472
9	*Bacteroides*	PMID:30090033
10	*Staphylococcus aureus*	PMID:7074582

TABLE 6 | Top 10 potential COPD-related microbes predicted by BWNMHMDA and 8 out of these 10 microbes have been confirmed by evidences.

Rank	Microbe	Evidence
1	*Tropheryma whipplei*	PMID:24460444
2	*Proteobacteria*	PMID:23071781
3	*Bacteroidetes*	PMID:29709671
4	*Prevotella*	PMID:30053882
5	*Clostridium difficile*	PMID:15655746
6	*Firmicutes*	PMID:24591822
7	*Helicobacter pylori*	PMID:15733502
8	*Lachnospiraceae*	Unconfirmed
9	*Staphylococcus aureus*	Unconfirmed
10	*Clostridia*	PMID:26852737

whipplei (Ranking first in the list of top 10 predicted microbes) has found in lower airway of human immunodeficiency virus-infected subjects (Segal et al., 2014; Sze et al., 2016). And also, it has been demonstrated that *Proteobacteria* (Ranking second in the list of top 10 predicted microbes) and *Firmicutes* (Ranking 3rd in the list of top 10 predicted microbes) will increase significantly with the development of COPD (Pragman et al., 2012). We published evidence for the top 10 potential COPD-related microbes predicted by BWNMHMDA in the **Table 6**.

Furthermore, in order to reconfirm the prediction performance of BWNMHMDA, we compared it with KATZHMDA in the case studies of these three kinds of same diseases, and as shown in the following **Table 7**, it is obvious that there are 10, 9, and 8 out of these top 10 microbes predicted by BWNMHMDA having been verified to be associated with the onset of asthma, colorectal carcinoma and COPD respectively, while there are only 4, 5, and 5 out of these top 10 microbes predicted by KATZHMDA having been verified to be associated with the onset of asthma, colorectal carcinoma, and COPD separately, which demonstrated that our prediction model BWNMHMDA could achieve better predictive hit rate in case above studies than the prediction model of KATZHMDA. And in addition, we published all these rankings of microbe-disease associations and top 10 disease-related microbes predicted by BWNMHMDA in **Supplementary Tables 1**, **2**, respectively, and hope that these data may provide some help to the future works of relevant researchers.

TABLE 7 | The number of of microbes having been confirmed by evidences in the top 10 potential disease-related microbes predicted by BWNMHMDA and KATZHMDA respectively in case studies of the three kinds of diseases such as Asthma, CRC, and COPD.

Model	Asthma	colorectal carcinoma	COPD
BWNMHMDA	10	9	8
KATZHMDA	4	5	5

6. DISCUSSION AND CONCLUSION

Human microbiome is normal flora for humans, which has been proved to be of symbiotic relationship with humans and harmless to humans. If the microbes that breed in the human body become "unhealthy," it will definitely affect the host's physical condition. People are continuing to explore the pathologic relationship between microorganisms and the human body through high-throughput sequencing technologies and analysis systems. However, it is a pity that their pathogenesis cannot be fully understood as yet. Considering that relying only on conventional experimental methods is time-consuming and laborious, in this article, we proposed a novel prediction model called BWNMHMDA to accelerate the process of inferring potential microbe-disease associations, in which, the core idea is to construct a weighted bidirectional microbe-disease association network and then convert it into a matrix for correlation probability calculation. While constructing the prediction model BWNMHMDA, we first downloaded known microbe-disease associations from the HDMDA database, and then, based on these downloaded associations, we constructed a heterogeneous network through adopting the Gaussian interaction profile kernel similarity to calculate the weights of nodes in the heterogeneous network. Moreover, based on the heterogeneous network, we further constructed a weighted bidirectional network by standardizing the weights of edges in the heterogeneous network and introducing a novel bidirectional recommendation method. Finally, we transformed the weighted bidirectional network into an integration matrix that can be utilized for prediction of potential microbe-disease associations. And simulation results show that BWNMHMDA can achieve reliable AUCs of 0.9127 and 0.8967 ± 0.0027 in the frameworks of LOOCV and 5-Fold CV respectively. And moreover, in the case studies of asthma, colorectal cancer, and COPD, there are 10, 9, and 8 out of the top 10 potential associated microbes predicted by BWNMHMDA having been verified by published literature evidence, which demonstrated that BWNMHMDA could provide valuable potential microbe-disease associations for future biological experiments. Certainly, there are some deficiencies in BWNMHMDA. For instance, there is a lack of negative samples in BWNMHMDA, and it may be possible to improve the predictive reliability of BWNMHMDA by identifying unrelated microbe-disease pairs. And moreover, in

BWNMHMDA, we adopt the Gaussian interaction profile kernel similarity to calculate the similarities between microbes, which may bias the similarity between some individual microbes. Hence, in subsequent work, we will introduce some effective methods such as Symptom-Based Disease Similarity (Zhou et al., 2014) to further improve the accuracy and efficiency of BWNMHMDA.

AUTHOR CONTRIBUTIONS

HL and LW conceptualized the study. HL and YW created the methodology, conducted the validation, and the data curation. HL, YW, HZ, and LW conducted the formal analysis. JJ, XF, HZ, and BZ oversaw the investigations. HL provided resources and prepared and wrote the original draft. LW wrote, reviewed and edited the manuscript, supervised the project, oversaw project administration, and acquired funding.

FUNDING

The project is partly sponsored by the National Natural Science Foundation of China (No.61873221, No. 61672447), the Natural Science Foundation of Hunan Province (No.2018JJ4058, No.2017JJ5036), and the CERNET Next Generation Internet Technology Innovation Project (No.NGII20160305, No.NGII20170109).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00676/full#supplementary-material

Supplementary Table 1 | Ranks of microbe-disease associations predicted by BWNMHMDA.

Supplementary Table 2 | Top 10 related microbes of all diseases.

REFERENCES

Astin, M., Griffin, T., Neal, R. D., Rose, P., and Hamilton, W. (2011). The diagnostic value of symptoms for colorectal cancer in primary care: a systematic review. *Br. J. Gen. Pract.* 61, e231–e243. doi: 10.3399/bjgp11X572427

Bhavsar, A. P., Guttman, J. A., and Finlay, B. B. (2007). Manipulation of host-cell pathways by bacterial pathogens. *Nature* 449, 827–834. doi: 10.1038/nature06247

Bibbins-Domingo, K., Grossman, D. C., Curry, S. J., Davidson, K. W., and Epling, J. W. E. A. (2008). Screening for colorectal cancer: U.S. preventive services task force recommendation statement. *Ann. Intern. Med.* 149:627. doi: 10.7326/0003-4819-149-9-200811040-00243

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

Busse, W. W. (2007). Expert panel report 3 (EPR-3): guidelines for the diagnosis and management of asthma–summary report 2007. *J. Allergy Clin. Immunol.* 120, S94–S138. doi: 10.1016/j.jaci.2007.09.029

Çalşkan, M., Bochkov, Y. A., Kreiner-Møller, E., Bønnelykke, K., Stein, M. M., Du, G., et al. (2013). Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *New Engl. J. Med.* 368, 1398–1407. doi: 10.1056/NEJMoa1211592

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., and Dewhirst, F. E. (2010). The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* 2010, baq013–baq013. doi: 10.1093/database/baq013

Chen, X. (2015). KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5:16840. doi: 10.1038/srep16840

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* 34, 1440–1440. doi: 10.1093/bioinformatics/btx773

Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. BioSyst.* 8:1970. doi: 10.1039/c2mb00002d

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016a). NLLSS: Predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput. Biol.* 12:e1004975. doi: 10.1371/journal.pcbi.1004975

Chen, X., Sun, Y.-Z., Guan, N.-N., Qu, J., Huang, Z.-A., Zhu, Z.-X., et al. (2018a). Computational models for lncRNA function prediction and functional similarity calculation. *Brief. Funct. Genomics* 18, 58–82. doi: 10.1093/bfgp/ely031

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018b). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.-H., and Liu, H. (2018c). BNPMDA: Bipartite network projection for MiRNA–disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2016b). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., and Yan, G.-Y. (2013). Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chen, X., and Yan, G.-Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.* 4:5501. doi: 10.1038/srep05501

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2013). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

Duncan, S. H., Belenguer, A., Holtrop, G., Johnstone, A. M., Flint, H. J., and Lobley, G. E. (2006). Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl. Environ. Microbiol.* 73, 1073–1078. doi: 10.1128/AEM.02340-06

Faith, J. J., Driscoll, M. E., Fusaro, V. A., Cosgrove, E. J., Hayete, B., Juhn, F. S., et al. (2007). Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* 36, D866–D870. doi: 10.1093/nar/gkm815

Fukugaiti, M. H., Ignacio, A., Fernandes, M. R., Júnior, U. R., Nakano, V., and Avila-Campos, M. J. (2015). High occurrence of fusobacterium nucleatum and *Clostridium difficile* in the intestinal microbiota of colorectal carcinoma patients. *Braz. J. Microbiol.* 46, 1135–1140. doi: 10.1590/S1517-838246420140665

Galiana, A., Aguirre, E., Rodriguez, J. C., Mira, A., Santibanez, M., Candela, I., et al. (2013). Sputum microbiota in moderate versus severe patients with COPD. *Eur. Respir. J.* 43, 1787–1790. doi: 10.1183/09031936.00191513

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

Gilbert, J. A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C. T., Brown, C. T., et al. (2010). Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. *Stand. Genomic Sci.* 3, 243–248. doi: 10.4056/sigs.1433550

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi: 10.1126/science.1124234

Gilstrap, D. L., and Kraft, M. (2013). Asthma and the host-microbe interaction. *J. Allergy Clin. Immunol.* 131, 1449–1450.e3. doi: 10.1016/j.jaci.2013.03.004

Guarner, F., and Malagelada, J.-R. (2003). Gut flora in health and disease. *Lancet* 361, 512–519. doi: 10.1016/S0140-6736(03)12489-0

Guilbert, T. W., Mauger, D. T., and Lemanske, R. F. (2014). Childhood asthma-predictive phenotype. *J. Allergy Clin. Immunol.* 2, 664–670. doi: 10.1016/j.jaip.2014.09.010

Hoppe, B., Groothoff, J. W., Hulton, S.-A., Cochat, P., Niaudet, P., Kemper, M. J., et al. (2011). Efficacy and safety of oxalobacter formigenes to reduce urinary oxalate in primary hyperoxaluria. *Nephrol. Dialys. Transpl.* 26, 3609–3615. doi: 10.1093/ndt/gfr107

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

Huang, Y. J. (2013). Asthma microbiome studies and the potential for new therapeutic strategies. *Curr. Allergy Asthma Rep.* 13, 453–461. doi: 10.1007/s11882-013-0355-y

Huang, Z.-A., Chen, X., Zhu, Z., Liu, H., Yan, G.-Y., You, Z.-H., et al. (2017). PBHMDA: Path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Jesmok, E. M., Hopkins, J. M., and Foran, D. R. (2016). Next-generation sequencing of the bacterial 16s rRNA gene for forensic soil comparison: a feasibility study. *J. Forens. Sci.* 61, 607–617. doi: 10.1111/1556-4029.13049

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika* 18, 39–43. doi: 10.1007/BF02289026

Ley, R. E., Backhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11070–11075. doi: 10.1073/pnas.0504978102

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

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2016). An analysis of human microbe–disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Mármol, I., de Diego, C. S., Dieste, A. P., Cerrada, E., and Yoldi, M. R. (2017). Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *Int. J. Mol. Sci.* 18:E197. doi: 10.3390/ijms18010197

Mason, M. R., Preshaw, P. M., Nagaraja, H. N., Dabdoub, S. M., Rahman, A., and Kumar, P. S. (2014). The subgingival microbiome of clinically healthy current and never smokers. *ISME J.* 9, 268–272. doi: 10.1038/ismej.2014.114

Mathers, C. D., and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.* 3:e442. doi: 10.1371/journal.pmed.0030442

Matsumoto, M., Sakamoto, M., Hayashi, H., and Benno, Y. (2005). Novel phylogenetic assignment database for terminal-restriction fragment length polymorphism analysis of human colonic microbiota. *J. Microbiol. Methods* 61, 305–319. doi: 10.1016/j.mimet.2004.12.009

Methé, 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

Mikaelyan, A., Köhler, T., Lampert, N., Rohland, J., Boga, H., Meuser, K., et al. (2015). Classifying the bacterial gut microbiota of termites and cockroaches: a curated phylogenetic reference database (DictDb). *Syst. Appl. Microbiol.* 38, 472–482. doi: 10.1016/j.syapm.2015.07.004

Moore, W., and Moore, L. H. (1995). Intestinal floras of populations that have a high risk of colon cancer. *Appl. Environ. Microbiol.* 61, 3202–3207.

Nagasaki, T., Matsumoto, H., Oguma, T., Ito, I., Inoue, H., Iwata, T., et al. (2017). Sensitization to *Staphylococcus aureus* enterotoxins in smokers with asthma. *Ann. Allergy Asthma Immunol.* 119, 408–414.e2. doi: 10.1016/j.anai.2017.08.001

Peng, L.-H., Yin, J., Zhou, L., Liu, M.-X., and Zhao, Y. (2018). Human microbe-disease association prediction based on adaptive boosting. *Front. Microbiol.* 9:2440. doi: 10.3389/fmicb.2018.02440

Pragman, A. A., Kim, H. B., Reilly, C. S., Wendt, C., and Isaacson, R. E. (2012). The lung microbiome in moderate and severe chronic obstructive pulmonary disease. *PLoS ONE* 7:e47305. doi: 10.1371/journal.pone.0047305

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. doi: 10.1038/nature11450

Quigley, E. M. (2013). Gut bacteria in health and disease. *Gastroenterol. Hepatol.* 9, 560–569. Available online at: https://scholar.google.com/scholar_lookup?title=Gut%20bacteria%20in%20health%20and%20disease&publication_year=2013&author=E.M.M.%20Quigley

Rauch, M. and Lynch, S. (2012). The potential for probiotic manipulation of the gastrointestinal microbiome. *Curr. Opin. Biotechnol.* 23, 192–201. doi: 10.1016/j.copbio.2011.11.004

Segal, L. N., Rom, W. N., and Weiden, M. D. (2014). Lung microbiome for clinicians. New discoveries about bugs in healthy and diseased lungs. *Ann. Am. Thorac. Soc.* 11, 108–116. doi: 10.1513/AnnalsATS.201310-339FR

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

Shah, P., Fritz, J. V., Glaab, E., Desai, M. S., Greenhalgh, K., Frachet, A., et al. (2016). A microfluidics-based *in vitro* model of the gastrointestinal human–microbe interface. *Nat. Commun.* 7:11535. doi: 10.1038/ncomms11535

Shen, X., Zhu, H., Jiang, X., Hu, X., and Yang, J. (2018). "A novel approach based on bi-random walk to predict microbe-disease associations," in *Intelligent Computing Methodologies*, eds D.-S. Huang, M. M. Gromiha, K. Han, and A. Hussain (Springer International Publishing), 746–752.

Shi, J.-Y., Huang, H., Zhang, Y.-N., Cao, J.-B., and Yiu, S.-M. (2018). BMCMDA: a novel model for predicting human microbe-disease associations via binary matrix completion. *BMC Bioinformatics* 19:281. doi: 10.1186/s12859-018-2274-3

Simpson, J. L., Daly, J., Baines, K. J., Yang, I. A., Upham, J. W., Reynolds, P. N., et al. (2015). Airway dysbiosis:haemophilus influenzaeandTropherymain poorly controlled asthma. *Eur. Respir. J.* 47, 792–800. doi: 10.1183/13993003.00405-2015

Sullivan, A., Hunt, E., MacSharry, J., and Murphy, D. M. (2016). The microbiome and the pathophysiology of asthma. *Respir. Res.* 17:163. doi: 10.1186/s12931-016-0479-4

Sze, M. A., Xu, S., Leung, J. M., Vucic, E. A., Shaipanich, T., Moghadam, A., et al. (2016). The bronchial epithelial cell bacterial microbiome and host response in patients infected with human immunodeficiency virus. *BMC Pulmon. Med.* 16:142. doi: 10.1186/s12890-016-0303-4

Thompson, C. C., Amaral, G. R., Campeão, M., Edwards, R. A., Polz, M. F., Dutilh, B. E., et al. (2014). Microbial taxonomy in the post-genomic era: rebuilding from scratch? *Arch. Microbiol.* 197, 359–370. doi: 10.1007/s00203-014-1071-2

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. doi: 10.1038/nature06244

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

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.e3. doi: 10.1016/j.jaci.2011.07.027

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., et al. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *Lancet* 388, 1545–1602. doi: 10.1016/S0140-6736(16)31678-6

Walker, A. W., Ince, J., Duncan, S. H., Webster, L. M., Holtrop, G., Ze, X., et al. (2010). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* 5, 220–230. doi: 10.1038/ismej.2010.118

Wang, F., Huang, Z.-A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Wen, L., Ley, R. E., Volchkov, P. Y., Stranges, P. B., Avanesyan, L., Stonebraker, A. C., et al. (2008). Innate immunity and intestinal

microbiota in the development of type 1 diabetes. *Nature* 455, 1109–1113. doi: 10.1038/nature07336

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344

Xuan, Z., Li, J., Yu, J., Feng, X., Zhao, B., and Wang, L. (2019). A probabilistic matrix factorization method for identifying lncRNA-disease associations. *Genes* 10:126. doi: 10.3390/genes10020126

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., et al. (2014). A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS ONE* 9:e87797. doi: 10.1371/journal.pone.0087797

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., et al. (2017). PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005455. doi: 10.1371/journal.pcbi.1005455

Yu, J., Ping, P., Wang, L., Kuang, L., Li, X., and Wu, Z. (2018). A novel probability model for LncRNA–disease association prediction based on the naïve bayesian classifier. *Genes* 9:345. doi: 10.3390/genes9070345

Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2365–2370. doi: 10.1073/pnas.0812600106

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

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms–disease network. *Nat. Commun.* 5:5212. doi: 10.1038/ncomms5212

Zhu, Q., Jin, Z., Wu, W., Gao, R., Guo, B., Gao, Z., et al. (2014). Analysis of the intestinal lumen microbiota in an animal model of colorectal cancer. *PLoS ONE* 9:e90849. doi: 10.1371/journal.pone.0090849

Conflict of Interest Statement: 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.

Copyright © 2019 Li, Wang, Jiang, Zhao, Feng, Zhao and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 18 April 2019
doi: 10.3389/fmicb.2019.00745

Bacterial Community Succession, Transmigration, and Differential Gene Transcription in a Controlled Vertebrate Decomposition Model

*Zachary M. Burcham[1], Jennifer L. Pechal[2], Carl J. Schmidt[3], Jeffrey L. Bose[4], Jason W. Rosch[5], M. Eric Benbow[2,6] and Heather R. Jordan[1]**

[1] Department of Biological Sciences, Mississippi State University, Starkville, MS, United States, [2] Department of Entomology, Michigan State University, East Lansing, MI, United States, [3] Department of Pathology, University of Michigan, Ann Arbor, MI, United States, [4] Department of Microbiology, Molecular Genetics, and Immunology, University of Kansas Medical Center, Kansas City, KS, United States, [5] Department of Infectious Disease, St. Jude Children's Research Hospital, Memphis, TN, United States, [6] Department of Osteopathic Medical Specialties, Michigan State University, East Lansing, MI, United States

Decomposing remains are a nutrient-rich ecosystem undergoing constant change due to cell breakdown and abiotic fluxes, such as pH level and oxygen availability. These environmental fluxes affect bacterial communities who respond in a predictive manner associated with the time since organismal death, or the postmortem interval (PMI). Profiles of microbial taxonomic turnover and transmigration are currently being studied in decomposition ecology, and in the field of forensic microbiology as indicators of the PMI. We monitored bacterial community structural and functional changes taking place during decomposition of the intestines, bone marrow, lungs, and heart in a highly controlled murine model. We found that organs presumed to be sterile during life are colonized by *Clostridium* during later decomposition as the fluids from internal organs begin to emulsify within the body cavity. During colonization of previously sterile sites, gene transcripts for multiple metabolism pathways were highly abundant, while transcripts associated with stress response and dormancy increased as decomposition progressed. We found our model strengthens known bacterial taxonomic succession data after host death. This study is one of the first to provide data of expressed bacterial community genes, alongside transmigration and structural changes of microbial species during laboratory controlled vertebrate decomposition. This is an important dataset for studying the effects of the environment on bacterial communities in an effort to determine which bacterial species and which bacterial functional pathways, such as amino acid metabolism, provide key changes during stages of decomposition that relate to the PMI. Finding unique PMI species or functions can be useful for determining time since death in forensic investigations.

Keywords: decomposition, postmortem microbiome, necrobiome, metatranscriptomic, metagenomic

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Jifeng Cai,
Central South University, China
Peter Anthony Noble,
University of Washington, United States

****Correspondence:***
Heather R. Jordan
jordan@biology.msstate.edu

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

Received: *17 October 2018*
Accepted: *25 March 2019*
Published: *18 April 2019*

Citation:
Burcham ZM, Pechal JL, Schmidt CJ, Bose JL, Rosch JW, Benbow ME and Jordan HR (2019) Bacterial Community Succession, Transmigration, and Differential Gene Transcription in a Controlled Vertebrate Decomposition Model. Front. Microbiol. 10:745. doi: 10.3389/fmicb.2019.00745

INTRODUCTION

Decomposing remains are a continuously shifting ecological system leading to changes in nutrient availability and microhabitat conditions, thus yielding a microbial consortium under constant selective pressures (Carter et al., 2007; Janaway et al., 2009; Hyde et al., 2013; Metcalf et al., 2016). Microorganisms associated with vertebrate remains are ubiquitous and may be deposited on a body

from the environment, invertebrate, or vertebrate scavengers of the necrobiome or were part of the existing microflora during life (Benbow et al., 2018). Vertebrate decomposition begins immediately after death and as tissue begins to breakdown during autolysis, an efflux of cellular components and nutrients occurs which is used by microbial, predominately bacterial, communities (Janaway et al., 2009; Hyde et al., 2013; Crippen et al., 2015). After an initial lag phase immediately following organismal death, bacterial communities begin to exponentially proliferate, transmigrate, and create specialized proteins that digest host tissues during putrefaction (Can et al., 2014; Pozhitkov et al., 2017). These metabolic changes drive the transformation of the environmental decomposing landscape through the release of waste products, nutrient depletion, oxygen availability, and pH cycles which further facilitates host tissue breakdown. In turn, bacterial communities involved in the decomposition process are highly dynamic and constantly competing for survival, nutrient acquisition, and habitat space.

Bacterial community succession occurs in a predictive manner associated with the time since death, or postmortem interval (PMI) (Finley et al., 2015; Hauther et al., 2015; Burcham et al., 2016). This discovery has led to extensive research of bacterial communities associated with remains for understanding intrinsic microbial ecology of decomposition, and more recently, to aid in biomarker discovery for forensic PMI estimation (Pechal et al., 2014; Javan et al., 2016; Metcalf et al., 2016). PMI estimation supports forensic investigations by providing a window of time when death occurred to help support or refute eyewitness accounts regarding events leading up to death. These community analyses have now been widely performed on animal models targeting the bacterial 16S gene for taxonomic community profiling and predictive function in an attempt to discover PMI-associated biomarkers (Damann and Carter, 2013; Metcalf et al., 2013, 2016; Pechal et al., 2014). Although research is also conducted on human remains, animal models provide investigators the ability for robust sample sizes in order to create statistically powerful studies, and control of habitat conditions, such as with insect colonization, temperature, etc. (Finley et al., 2015; Hyde et al., 2017; Pechal et al., 2018). In a study using terrestrial swine models, microbial diversity decreased as decomposition progressed with Proteobacteria being the dominant phylum in early stages, while Firmicutes dominated late stages. In those studies, Clostridiaceae was one of the most dominant families toward the end of decomposition (Pechal et al., 2014). Additionally, in terrestrial murine models, during bloat stage the abdominal cavity has increased relative abundances of anaerobic gut microbiota, Lactobacillaceae, and Bacteroidaceae, but transitions to contain more oxygen-tolerant (i.e., Enterobacteriaceae) bacteria following burst (Metcalf et al., 2013, 2016). Overall, animal models have shown similarities to microbial succession discovered in human remains such as the transition from aerobic to anaerobic bacteria, prevalence of *Clostridium* in late stages, and bacterial community differences based on body location (Hyde et al., 2013; Finley et al., 2015; Hauther et al., 2015). These data are promising for the use of animal decomposition models as surrogates for microbial involvement within human remains, and for measuring correlates of microbial structure and function during decomposition.

While recognizing the broad trends in microbial contributions to decomposition is important for narrowing research foci, it is imperative that we study microbial interactions at a finer resolution with reference baseline activity, if specific, usable biomarkers are going to be detected. This fine resolution along with baseline data approach is important for teasing apart bacterial taxonomic and functional succession variability in the decomposition process, which may affect potential microbial biomarker discovery. This approach raises the call for highly controlled studies with the ability to focus on the microbial interactions solely associated with the host so that a host baseline can be established and built upon with other variables (i.e., climate, soil, and scavengers) and external environmental microbial communities.

Our study aims to develop the host postmortem microbial structure in conjunction with functional activity using a highly controlled laboratory setting, without the introduction of external environmental factors. We are also among the first to utilize postmortem microbial metatranscriptomic analyses. While metagenomic investigation provides insights about microbial community structure and functional potential, metatranscriptomic investigation is a useful tool to shed light on the active functional profiles of a microbial community. Together, both analytic methods provide knowledge on both the microbial diversity as well as genes actively involved in ecosystem processes. For instance, are postmortem microbiome changes observed and measured during decomposition succession reflective of coordinated responses, plasticity within an individual microbial species, or consequences of environment disturbance? Data have shown that many coexisting but taxonomically distinct microbes can encode genes for the same metabolic functions, which may blur the association between community composition and ecosystem functioning (Louca et al., 2018). However, for functions performed by only a few taxa, the sensitivity and resilience of this function may closely follow changes in the abundance of those taxa (Langenheder et al., 2006; Delgado-Baquerizo et al., 2016; Louca et al., 2018). Understanding mechanisms by which microbial functions vary, including shifts in community composition, gene expression patterns, or density, will benefit studies in decomposition ecology and forensic science, in order to determine what changes are most meaningful during decomposition, and will aid in biomarker discovery for PMI estimation.

MATERIALS AND METHODS

S. aureus/*C. perfringens* Preparation and Murine Inoculation

A detailed description on the construction of *Staphylococcus aureus* KUB7 and *Clostridium perfringens* inoculums along with the murine inoculation, sacrifice, and organ harvest can be found in Burcham et al. (2016). An experimental procedure flow chart is detailed in **Figure 1**. Briefly, *S. aureus* KUB7 is a constructed strain that constitutively expresses a red fluorescent protein.

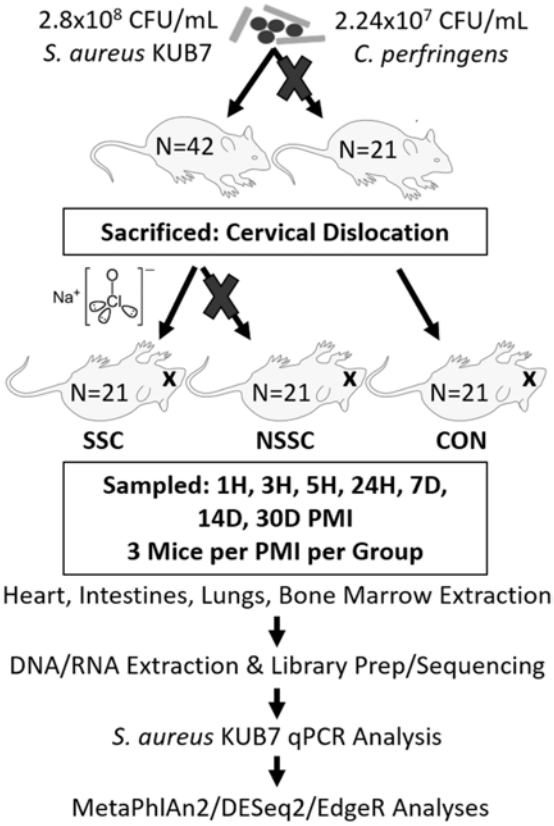


FIGURE 1 | Experimental flow chart. A visual representation of the experimental design and downstream analyses. Mouse groups are represented as SSC = surface sterilized, colonized; NSSC = nonsurface sterilized, colonized; and CON = control.

C. perfringens type A strain WAL 14572 is non-fluorescent and was obtained from ATCC. *S. aureus* KUB7 and *C. perfringens* were grown to exponential phase in 100 mL tryptic soy broth (TSB) or reinforced clostridial medium (RCM), respectively. Sixty-four 1 mL cultures obtained from the original culture were pelleted and *S. aureus* KUB7 was resuspended in 7 μL TSB while *C. perfringens* was resuspended in 7 μL of RCM supplemented with 60 g/L sucrose. The resuspensions were combined and used for nasal inoculation through inhalation in 42 isoflurane sedated SKH-1 female mice (N = 42) obtained from Charles Rivers Laboratories with 21 mice not being inoculated as controls (N = 21) for a total of 90 mice (N = 63). The final inhalation inoculum was 2.8×10^8 CFU/mL and 2.24×10^7 CFU/mL for *S. aureus* KUB7 and *C. perfringens*, respectively. These two species have been shown to colonize living humans, albeit in separate niches, and have been shown to transmigrate and produce enzymes that break down tissues during decomposition (Kellerman et al., 1976; Melvin et al., 1984; Tuomisto et al., 2013; Burcham et al., 2016). We introduced chromosomally labeled red fluorescent *S. aureus* and non-labeled *C. perfringens* to the nasal pharynx and upper respiratory tract of mice by inhalation. This location was selected as it is a natural colonization site for *S. aureus* in humans and *C. perfringens* should not thrive in these oxygen rich environments. All animal experiments were conducted according to Mississippi State University IACUC approved protocol 14–102.

Murine Sacrifice, Controlled Decomposition, and Organ Harvest

Twenty-four hours after inoculation, all mice were sacrificed by cervical dislocation, as previously described (Burcham et al., 2016). Twenty-one of the 42 inoculated mice were randomly chosen to be surface sterilized to disrupt the skin microbial communities. The surface sterilized mice were submerged up to below the mouth in a 10% bleach solution for 45 s avoiding sterilization of the mouth, nares, and ears to prevent the bleach solution from entering the body. The bleach solution was rinsed twice successively with distilled water. All 63 mice were individually placed in a Nalgene bottle top 0.2 μm filter container (Thermo Scientific) and sealed with Parafilm M to prevent environmental microbial and insect contamination. Mouse carcasses were allowed to naturally decompose within a bilaminar flow hood at ambient room temperature for up to 30 d during July 2015 in Starkville, MS, United States.

Three mice per treatment (control, inoculated with no surface sterilization, and inoculated with surface sterilization) were analyzed per time point (T = 1 h, 3 h, 5 h, 24 h, 7 d, 14 d, and 30 d postmortem). All tissue harvesting, and subsequent analyses were performed in a bilaminar flow hood under BSL2 and sterile conditions. A sterile scalpel blade was used to cut through the right hind leg and femur. The separated leg was used to obtain bone marrow from the femur using a sterile syringe containing nuclease-free water. A second sterile scalpel blade was used to dissect each mouse from the ventral side to remove the lungs, heart, and composite of the intestines using sterile forceps, individual for each organ. The bone marrow solution and each organ were transferred to individual 2.0 mL screw cap tubes. Each organ was crushed with a sterile cotton swab excluding the bone marrow solution. All organ swabs were swabbed on specialized agar plates for plate count determination discussed in Burcham et al. (2016). Afterward, RNAlater® (Ambion) was added to each tube and the samples were stored in −20°C until nucleic acid extraction. Due to the variation of decomposition across individuals, some organs were no longer discernable in the later stages of decomposition and composite samples of the organ location were collected. Decomposition stages present at each postmortem timepoint are represented in **Figure 2**.

Nucleic Acid Extraction and Purification

The following postmortem timepoints were chosen for sequencing to focus on the early decomposition processes: 1 h, 3 h, 5 h, 24 h, and 7 d. Therefore, to obtain both RNA and DNA from these samples, the dual extraction method using the TRIzol™ Reagent (Thermo Fisher) standard issued protocol was performed on the preserved samples after being spun down and RNAlater® removed. Briefly, 100 mg of tissue or the pelleted bone marrow was added to 1 mL of TRIzol™ Reagent and a mix of 0.1/0.5 mm glass beads. The samples were homogenized in a bead beater with phase separation following chloroform

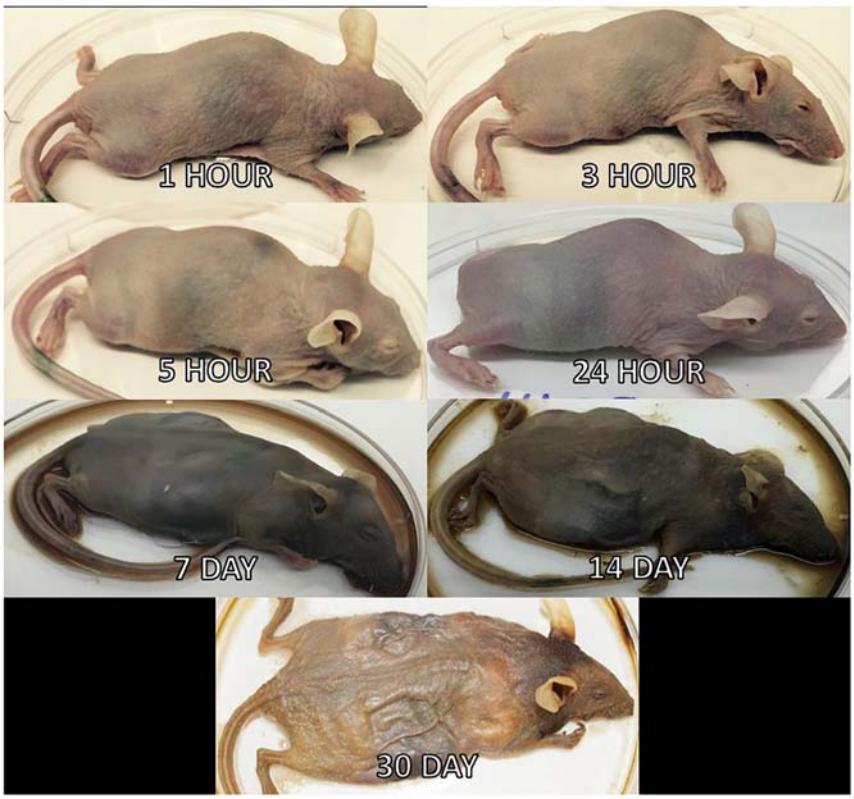


FIGURE 2 | Mouse decomposition stages. Images of mice before dissection and organ harvest for their selected postmortem times to represent the decomposition stages present at each timepoint.

addition. RNA was precipitated with isopropanol, washed with 75% ethanol, and resuspended in 50 μL of nuclease-free water, and incubated at 60°C for 15 min. The samples were purified using the PowerClean® Pro RNA Clean-Up Kit (Qiagen), quantified fluorometrically using a Qubit 2.0™ (Invitrogen), and then stored at −80°C. DNA was extracted during the TRIzol™ Reagent protocol as described above. DNA within the sample interphase was precipitated with ethanol, washed with 0.1 M sodium citrate in 10% ethanol, washed with 75% ethanol, and resuspended in 0.6 mL of 8 mM NaOH. The samples were purified using a PowerClean® Pro DNA Clean-Up Kit (Qiagen). All samples were quantified fluorometrically using a Qubit 2.0™ (Invitrogen) and stored at −20°C. The non-sequenced timepoints (T = 14 d and 30 d) had DNA extracted using a modified protocol of that discussed in Williamson et al. (2014).

C. perfringens WAL 14572 Detection

Difficulty in creating a fluorescently tagged *C. perfringens* strain resulted in detection of *C. perfringens* only in association with bacterial community analysis by metagenomic sequencing. Metagenomic sequencing analysis allowed determination of whether *C. perfringens* introduced nasally caused increased *C. perfringens* levels during early decomposition, particularly in less microbially rich organs, when compared to non-inoculated mice.

S. aureus KUB7 Quantitative PCR Analysis

Transmigration of *S. aureus* KUB7 through body tissues as decomposition progressed was tracked at all postmortem times (1 h, 3 h, 5 h, 24 h, 7 d, 14 d, and 30 d) using qPCR by amplifying the red fluorescent protein gene incorporated in the genome of *S. aureus* KUB7. A 25 μL reaction was created using 3 μL template, 12.5 μL SsoAdvanced™ Universal Probes Supermix (Bio-Rad), 1 μL forward primer (5′-TTGAAGGTGAAGGTGAAGGA-3′), 1 μL reverse primer (5′-TGCAAATGGTAATGGACCAC-3′), 2.5 μL FAM probe (5′-6FAMTGGAAGGTACACAAACAGCAAAAMGBNFQ-3′), and 5 μL nuclease-free water. The reaction was amplified using a Bio-Rad C1000 Touch™ thermal cycler and measured with a Bio-Rad CFX96™ real-time system with the following cycling conditions: 95°C for 10 min and (95°C for 15 s, 56°C for 30 s) × 40 cycles. Each sample was analyzed in duplicate to obtain an average CQ value and copies/run, which was extrapolated to obtain the mean genomic units of KUB7 per sample. The R statistical package was used to remove outliers in each organ as determined by Cook's distance (R Core Team, 2013). The mean genomic units were log transformed, the mean log genomic units (logGU) ± standard error of the mean for each postmortem time for each organ were calculated. Wilcoxon rank sum significance testing between (non)surface sterilization did

not show a difference between treatments (intestines: $p = 0.79$, bone marrow: $p = 0.56$, heart: $p = 0.16$, lungs: $p = 0.84$). Therefore, both treatments were combined and treated as the same for further testing. Significant differences between postmortem times were tested using a Kruskal–Wallis rank sum test for each organ. Significance was based on a $p < 0.05$.

Metagenomic/Metatranscriptomic Library Preparation

The following postmortem timepoints were chosen for sequencing to focus on the early decomposition processes: 1 h, 3 h, 5 h, 24 h, and 7 d. Total DNA libraries were created using the NEBNext® Ultra™ DNA Library Prep Kit and NEBNext® Multiplex Oligos (Dual Index Primers) for Illumina® (New England BioLabs) protocols on all samples. Total RNA libraries were created using the NEBNext® Ultra™ RNA Library Prep Kit and NEBNext® Multiplex Oligos (Dual Index Primers) for Illumina® (New England BioLabs) protocols for use with purified mRNA or rRNA depleted RNA on the nonsurface sterilized and control samples excluding the lungs since they did not sequence well for metagenomic analysis. These protocols were chosen to maintain rRNA and mRNA within the RNA sample in order to preserve 16S rRNA genes and mRNA for both structural and differential transcript expression analysis. DNA samples were used for metagenomics community analysis.

Whole Metagenome/Metatranscriptome Shotgun Sequencing and Processing

High-throughput whole metagenome and metatranscriptome sequencing was performed by St. Jude Children's Research Hospital on an Illumina HiSeq2000 with 2 × 100 bp paired end (PE) read lengths. Sequences were initially trimmed by the sequencing facility using TrimGlare v0.4.2 (Krueger, 2015), but a more stringent quality trimming was performed using Trimmomatic v0.33 (Bolger et al., 2014). Metagenome sequences were trimmed to remove nucleotides in a four-position sliding window with an average phred33 score less than 28, and read lengths less than 36 bp. The trimmed metagenomic sequences were then used for bacterial community analysis. Metatranscriptome sequences were input in the SAMSA2 pipeline (Westreich et al., 2018). SAMSA2 was used to merge the paired-end sequences with PEAR v0.9.10 (Zhang et al., 2014), and then trimmed in a four-position sliding window with an average phred33 score less than 20, and read lengths less than 99 bp with Trimmomatic v0.3 (Bolger et al., 2014). SortMeRNA v2.1 (Kopylova et al., 2012) was used to remove bacteria/archaea 16S/23S rRNA genes and eukaryotic 18S/28S/5S/5.8S rRNA genes based on the documentation recommended SILVA and Rfam databases (Quast et al., 2013; Yilmaz et al., 2014; Kalvari et al., 2018). Sample identifiers and metadata can be found in **Table 1**.

Metagenomic Bacterial Community Analysis

Relative abundance of the bacterial genera present in each metagenomic sample was determined using MetaPhlAn2 (Truong et al., 2015). MetaPhlAn2 uses roughly 1 million clade-specific markers from over 7500 species to characterize the microbial taxonomic profiles. Genera that constituted less than 3% of sample were grouped as rare taxa to reduce sampling noise, but were not grouped as rare taxa for community metrics so that rare taxa could be account for between test groups. The relative abundances were used to determine the log genera richness, Shannon diversity indices, and Bray–Curtis and binary Jaccard distance indices using the R statistical package vegan v2.5-1 (Oksanen et al., 2018). The metadata factors (colonization, sterilization, organ, and postmortem time) were used in a type-II Multivariate Analysis of Variance (MANOVA) additive model to test for differences in log genera richness, Shannon diversity indices, and Pielou's evenness using the R statistical package car v3.0-0 (Fox and Weisbert, 2011). The Bray–Curtis and binary Jaccard indices were tested against the metadata factors in a type-II permutational MANOVA additive model using the R statistical package RVAideMemoire v0.9-69-3 (Herve, 2018). Both distance indices were calculated based on their nature to account for taxonomic abundances (Bray–Curtis) or to treat taxonomic data as presence–absence (Jaccard). Analysis of both distances is important as presence–absence data give more statistical weight to rare taxa while using abundances give more statistical weight to the taxa of higher abundance. Determination of the distance-based redundancy analysis (dbRDA) explanatory variables (metadata factors) was performed using forward selection with both distance indices after the recommended "method 1" transformation by Legendre and Anderson (1999). Forward selection is a method of stepwise regression which starts with an empty model and then adds the metadata factor which improves the model the most. This metadata factor addition continues with the remaining metadata until no more factors significantly improve the model. The forward selection determined significant explanatory variables were used to create dbRDA ordination plots with both distance indices and the significant explanatory variables as interactions. The genera driving the ordination distances were determined with permutation environmental fit by their p-value ($p < 0.05$) and R^2. The dbRDA and environmental fit analyses were performed using the R statistical package vegan v2.5-1 (Oksanen et al., 2018).

Transcript Annotation and Differential Expression Analysis

Ribosomal RNA depleted RNA transcripts were used in the SAMSA2 pipeline (Westreich et al., 2018). The reads were annotated using the DIAMOND sequence aligner to a database created from the March 2018 NCBI nr-protein database (Buchfink et al., 2015). The best protein hit for each read in the sample was aggregated for differential expression analysis using the R statistical packages edgeR v3.22.1 and DESeq2 v1.20.0 (Robinson et al., 2010; McCarthy et al., 2012; Love et al., 2014). Treatments were not differentiated, as community analysis showed no difference between colonization nor surface sterilization. In edgeR, the counts per million were computed and transcripts that were not present more than 1 CPM in at least two samples were removed to reduce noise. In DESeq2,

TABLE 1 | Sample metadata.

Sample	Treatment	Colonized	Sterilized	PMI	Organ	Number of DNA reads	Number of RNA reads	Genera richness
MG01	C	N	N	1H	BM	519	10642207	0
MG02	C	N	N	1H	HT	141	9424377	0
MG03	C	N	N	1H	INT	6869462	11198249	5
MG04	C	N	N	1H	LU	246592	NA	0
MG05	NS	Y	N	1H	BM	12541054	6174552	4
MG06	NS	Y	N	1H	HT	4069	4370991	0
MG07	NS	Y	N	1H	INT	13588235	9345718	9
MG08	NS	Y	N	1H	LU	7921779	NA	0
MG09	S	Y	Y	1H	BM	12371788	NA	2
MG10	S	Y	Y	1H	HT	114640	NA	0
MG11	S	Y	Y	1H	INT	23356918	NA	14
MG12	S	Y	Y	1H	LU	5764224	NA	0
MG13	C	N	N	3H	BM	3085317	6966014	0
MG14	C	N	N	3H	HT	68	6198445	0
MG15	C	N	N	3H	INT	1519870	10537046	3
MG16	C	N	N	3H	LU	552826	NA	0
MG17	NS	Y	N	3H	BM	9372874	6585375	2
MG18	NS	Y	N	3H	HT	11238	8625237	0
MG19	NS	Y	N	3H	INT	3748353	9724755	6
MG20	NS	Y	N	3H	LU	16267397	NA	0
MG21	S	Y	Y	3H	BM	12034029	NA	0
MG22	S	Y	Y	3H	HT	10885	NA	0
MG23	S	Y	Y	3H	INT	3403829	NA	3
MG24	S	Y	Y	3H	LU	13384061	NA	0
MG25	C	N	N	5H	BM	2082085	6939217	0
MG26	C	N	N	5H	HT	671145	7417681	0
MG27	C	N	N	5H	INT	12924794	9505797	6
MG28	C	N	N	5H	LU	667927	NA	0
MG29	NS	Y	N	5H	BM	14411164	6502696	3
MG30	NS	Y	N	5H	HT	1353231	6454480	1
MG31	NS	Y	N	5H	INT	4597616	5133461	7
MG32	NS	Y	N	5H	LU	12237082	NA	0
MG33	S	Y	Y	5H	BM	12515225	NA	0
MG34	S	Y	Y	5H	HT	14546830	NA	12
MG35	S	Y	Y	5H	INT	17560686	NA	10
MG36	S	Y	Y	5H	LU	20161875	NA	1
MG37	C	N	N	24H	BM	5704975	3622308	0
MG38	C	N	N	24H	HT	1583416	3309305	0
MG39	C	N	N	24H	INT	9105711	5814593	2
MG40	C	N	N	24H	LU	654955	NA	0
MG41	NS	Y	N	24H	BM	5137625	1608180	0
MG42	NS	Y	N	24H	HT	5126506	5397803	0
MG43	NS	Y	N	24H	INT	6336406	771048	3
MG44	NS	Y	N	24H	LU	11441973	NA	0
MG45	S	Y	Y	24H	BM	13481461	NA	0
MG46	S	Y	Y	24H	HT	4407690	NA	0
MG47	S	Y	Y	24H	INT	34742830	NA	4
MG48	S	Y	Y	24H	LU	20444728	NA	0
MG49	C	N	N	7D	BM	59584111	778479	2
MG50	C	N	N	7D	HT	7277358	3818602	3
MG51	C	N	N	7D	INT	142284	2124152	3
MG52	C	N	N	7D	LU	16679	NA	0

(Continued)

TABLE 1 | Continued

Sample	Treatment	Colonized	Sterilized	PMI	Organ	Number of DNA reads	Number of RNA reads	Genera richness
MG53	NS	Y	N	7D	BM	63145271	537412	2
MG54	NS	Y	N	7D	HT	10769919	3795345	1
MG55	NS	Y	N	7D	INT	6286983	5305095	10
MG56	NS	Y	N	7D	LU	11484731	NA	0
MG57	S	Y	Y	7D	BM	16535712	NA	1
MG58	S	Y	Y	7D	HT	9303940	NA	1
MG59	S	Y	Y	7D	INT	27283847	NA	21
MG60	S	Y	Y	7D	LU	13993375	NA	3

Table demonstrates the sample identification with their treatment (C, control; NS, nonsurface sterilized; S, sterilized), if they were colonized (Y, yes; N, no), if they were sterilized (Y, yes; N, no), their PMI (1 h, 3 h, 5 h, 24 h, and 7 d), the organ the sample was obtained from (BM, bone marrow; HT, heart; INT, intestines; LU, lungs), the average number of DNA reads between paired-end reads, the number of RNA reads after paired-end merging, and the genera richness. Read counts are after quality control.

transcripts that were not present in at least two samples with counts above 1 were removed. Normalization by library size was performed based on each packages' recommended method. Normalization based on library size allows for samples with different numbers of RNA reads to be compared against each other without the read count differences affecting the results. This was especially important in our study as we found that our samples tended to decrease in RNA read abundance as decomposition progressed. EdgeR dispersion parameters for the negative binomial model estimations were determined individually per organ with bone marrow using bin-spline, heart using power, and intestines using spline. These dispersion methods were chosen independently for each organ based on which method provided a trendline which most fits the variation distribution. DESeq2 dispersions were estimated based on the default settings from the DESeq command. All models were created with no *y*-intercept and postmortem timepoint groups as the explanatory variable for each organ. EdgeR models were negative binomial generalized log-linear models (GLM) with quasi-likelihood and DESeq2 models were negative binomial GLMs with likelihood ratio tests between the full and reduced model without timepoints. For both methods, differential expression was tested to contrast the timepoint groups (early vs. middle, middle vs. late, early vs. late) for each organ and significantly differentially expressed transcripts were identified based on a Benjamini–Hochberg corrected p-value < 0.10 above 1 log2-fold-change threshold. Only transcripts determined to be significant by both methods were considered truly significantly expressed between the compared groups, to reduce procedural bias. Transcripts were annotated into pathways based on the KEGG and UniProt databases (Kanehisa and Goto, 2000; The UniProt Consortium, 2017).

RESULTS

S. aureus KUB7 Quantitative PCR

The *S. aureus* KUB7 mean log genomic units (logGU ± SE) in the lungs decreased from 6.91 ± 2.32 logGU to 4.21 ± 2.74 logGU then to 0 ± 0 logGU from 1 to 3 then 5 h after death, respectively. After 5 h, the mean logGU increased from 0 ± 0 logGU to 3.95 ± 1.90 logGU after 24 h, 18.97 ± 2.69 logGU after 7 d, and reached its maximum concentration (19.09 ± 6.77 logGU) after 14 d. Finally, the mean logGU decreased to levels similar to early timepoints of 6.75 ± 2.17 logGU at 30 d postmortem (**Figure 3A**). Overall, the lungs began with *S. aureus* KUB7 present then diminished within 5 h postmortem. Afterward, *S. aureus* KUB7 increased rapidly until returning to low levels after 14 d postmortem.

The *S. aureus* KUB7 mean logGU ± SE in the intestines stayed relatively consistent the first three timepoints (T = 1 h, 3 h, and 5 h) with logGU of 11.33 ± 0.95, 13.90 ± 0.58, and 12.40 ± 0.87, respectively. A decrease occurred at 24 h to 7.57 ± 3.40 logGU. The logGU increased to its maximum concentration after 7 d to 18.08 ± 2.0 logGU. *S. aureus* KUB7 detection decreased to below starting levels after 14 d (4.85 ± 1.57 logGU) to the minimum (2.86 ± 1.81 logGU) after 30 d postmortem time (**Figure 3B**). Overall, *S. aureus* KUB7 concentrations in the intestines during early postmortem times were relatively stable until exponential growth after 7 d, but then immediately began to decrease to levels below initial sampling.

Staphylococcus aureus KUB7 detection in the heart remained at zero until 7 d postmortem (11.72 ± 2.64 logGU) and increased to its maximum (18.49 ± 7.25 logGU) after 14 d. After 30 d, detection decreased to levels near zero (3.07 ± 3.07 logGU) (**Figure 3C**). Overall, the heart did not appear to be colonized, based on qPCR detection, by *S. aureus* KUB7 until 7 d postmortem, but once established, grew substantially until decreasing back to almost undetectable levels by 30 d postmortem.

Staphylococcus aureus KUB7 detection in the bone marrow started at 6.08 ± 3.13 logGU and decreased to 0 ± 0 logGU after 3 h. Detection increased to 1.88 ± 1.88 logGU after 5 h to 4.17 ± 2.64 logGU after 24 h and then to the maximum of 8.18 ± 3.71 logGU at 7 d. Detection decreased to 4.35 ± 3.27 logGU after 14 d then to 0 ± 0 logGU by 30 d postmortem (**Figure 3D**). Surprisingly, *S. aureus* KUB7 was detected almost immediately in the bone marrow after death, but then quickly decreased after 3 h until steadily increasing to its highest concentration after 7 d, then dissipating.

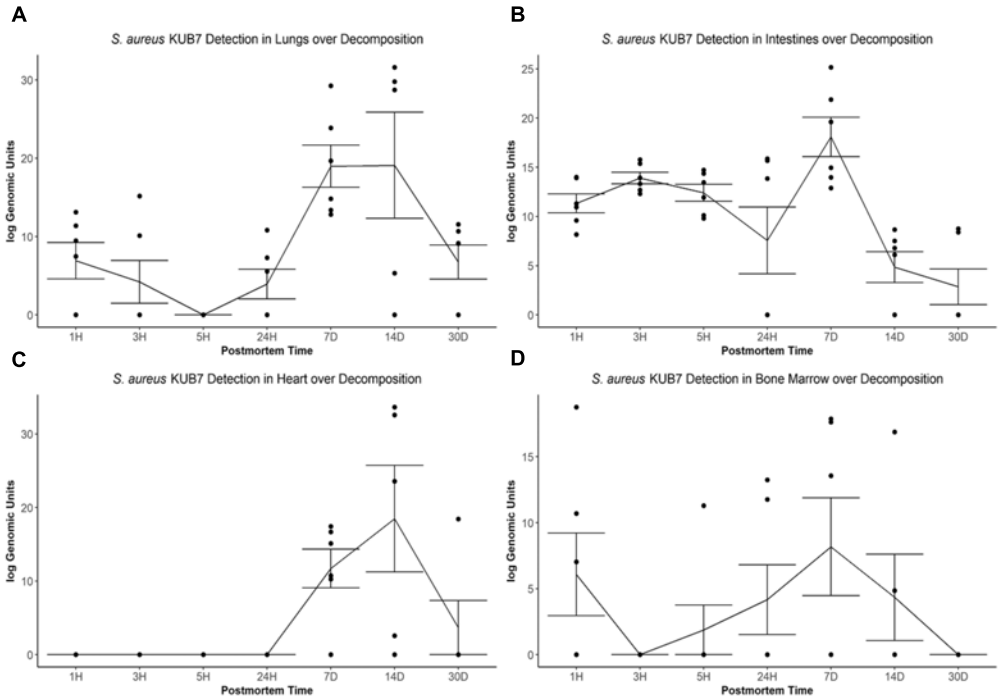


FIGURE 3 | *S. aureus* KUB7 detection with qPCR in organs as decomposition progresses. The log genomic units are plotted on the *y*-axis and postmortem time on the *x*-axis. Each circle represents a sample and the line represents the mean (±SE) log genomic units of *S. aureus* KUB7 during decomposition in the **(A)** lungs, **(B)** intestines, **(C)** heart, and **(D)** bone marrow.

When testing for a significance difference between postmortem times in each organ using Kruskal–Wallis rank sum test based on logGU we were able to reject the null hypothesis that the mean logGU of all timepoints are equal in the lungs (χ^2 = 19.71, *df* = 6, *p* = 0.003), intestines (χ^2 = 23.94, *df* = 6, *p* = 0.0005), and heart (χ^2 = 25.28, *df* = 6, *p* = 0.0003). We were not able to reject the null hypothesis for the bone marrow samples.

Metagenomic Bacterial Genera Relative Abundance

Twenty-six unique genera were detected across the 60 samples for a total of 144 detections (min = 0, max = 21, mean = 2.4, *SD* = 4.08). Thirty-one samples did not match classified bacteria. These samples were pre-dominantly associated with early–middle (≤24 h) postmortem times in the lungs, bone marrow, and heart. In the lungs, only two samples provided community profiles (**Figure 4A**). Sample MG36 contained 100% *Lactobacillus* at 5 h and sample MG60 was made up of approximately 44% and 55% *Clostridium* and *Staphylococcus* at 7 d, respectively. As expected, the intestines provided the most robust abundance community profiles (**Figure 4B**). Early (≤5 h) postmortem times in the intestines showed dominating bacterial genera consisting of *Parabacteroides* (μ = 47.0%), *Mucispirillum* (μ = 29.6%), and *Lactobacillus* (μ = 14.4%). At 24 h, there was a decrease of relative abundance in *Parabacteroides* (μ = 10.5%), disappearance of *Mucispirillum* and increase of *Lactobacillus* (μ = 86.3%). At 7 d, *Lactobacillus* (μ = 30.6%) had decreased, allowing for the increase of *Anaerostipes* (μ = 28.6%), *Clostridium* (μ = 16.1%), and *Enterococcus* (μ = 13.3%). Bacterial genera within the heart were only detected at 5 h and 7 d (**Figure 4C**). At 5 h, sample MG30 showed 100% *Escherichia*, while MG34 showed a diverse community with the highest percentage of genera detected being *Candidatus Arthromitus* (31.7%), *Parabacteroides* (24%), *Anaerostipes* (19.3%), and *Dorea* (10.7%). At 7 d, the highest percentage of genera detected was *Clostridium* (μ = 72.1%), with *Lactobacillus* (μ = 15.5%) and Peptostreptococcaceae spp. (μ = 12.4%) being detected in one sample. In the bone marrow, four out of nine samples in early (≤24 h) postmortem times provided detected genera (**Figure 4D**). The early time group genera detected were Propionibacteriaceae spp. (μ = 10.6%), *Staphylococcus* (μ = 9.1%), *Propionibacterium* (μ = 8.8%), *Enterococcus* (μ = 8.3%), and *Pseudomonas* (μ = 7.1%). At 7 d, similar to heart sequences, Clostridium dominated the samples (μ = 84.0%), with Peptostreptococcaceae spp. (μ = 11.7%) and Pseudomonas (μ = 4.3%) also being detected.

Metagenomic Bacterial Community Analyses

Type-II MANOVA additive model testing for log genera richness only showed a significant difference among organs (*SS* = 8.05, *df* = 2, *F* = 9.28, *p* = 0.002) with pairwise analysis determining differences between intestines–bone marrow (*p* = 0.01) and

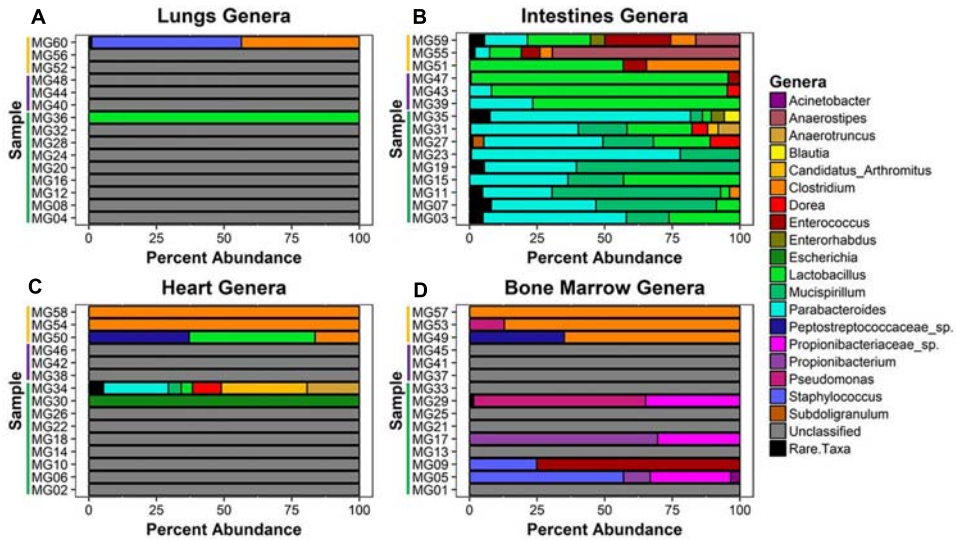


FIGURE 4 | Bacterial genera relative abundance in samples at postmortem times. The bacteria genera detected in **(A)** lungs, **(B)** intestines, **(C)** heart, and **(D)** bone marrow samples are represented by their relative percent abundance on the *x*-axis with the sample on the *y*-axis. The samples are labeled by postmortem timepoint group with the color bars [green = early (1 h, 3 h, 5 h), purple = middle (24 h), and yellow = late (7 d)]. Genera that constituted less than 3% of sample were grouped as rare taxa to reduce sampling noise. Relative abundances were determined using MetaPhlAn v2.0.

intestines–heart (p = 0.02). Type-II MANOVA additive model testing Shannon diversity indices only showed a significant difference among organs (SS = 1.71, df = 2, F = 4.07, p = 0.04) with pairwise analysis determining no significant pairs. Type-II MANOVA additive model testing squared Simpson's diversity indices showed no difference. The type-II permutational MANOVA additive model for Bray–Curtis distance showed a difference between organs (SS = 1.74, mean sq. = 0.87, df = 2, F = 4.68, p = 0.001) and postmortem times (SS = 2.81, mean sq. = 0.70, df = 4, F = 3.79, p = 0.001). The type-II permutational MANOVA additive model for binary Jaccard distance showed a difference among organs (SS = 1.83, mean sq. = 0.92, df = 2, F = 4.19, p = 0.001) and postmortem times (SS = 2.2, mean sq. = 0.55, df = 4, F = 2.51, p = 0.001). It is important to note that bacterial diversity analyses comparing the treatments (control, inoculated with no surface sterilization, and inoculated with surface sterilization) showed no differentiation. We detected no difference between mice bacterial communities that were or were not surface sterilized nor did we detect a difference between the mice bacteria communities that were or were not inoculated with *S. aureus* and *C. perfringens*. Because of this we were able to combine the data across treatments to obtain higher sample sizes for comparing community structure and function across postmortem times and organs.

The forward selection model for all metadata factors using dbRDA ordination on Bray–Curtis distances determined the explanatory variables to be PMI (R^2 = 0.18, df = 4, AIC = 69.90, F = 2.45, p = 0.002) and organ (R^2 = 0.32, df = 2, AIC = 66.32, F = 3.24, p = 0.002). These explanatory variables were treated as interactions to create the Bray–Curtis dbRDA ordination (**Figure 5A**). The environmental fit of the organ (R^2 = 0.48, p = 0.001) and PMI (R^2 = 0.46, p = 0.003) variables were significant with improved fits. Seven genera were identified to be structuring the ordination by environmental fit (**Figure 5A**). The forward selection model for all metadata factors using dbRDA ordination on Jaccard distances determined the explanatory variables to be PMI (R^2 = 0.21, df = 4, AIC = 62.18, F = 2.76, p = 0.002) and organ (R^2 = 0.35, df = 2, AIC = 58.43, F = 3.33, p = 0.002). These explanatory variables were treated as interactions to create the Jaccard dbRDA ordination (**Figure 5B**). The environmental fit of the organ (R^2 = 0.32, p = 0.001) and PMI (R^2 = 0.54, p = 0.001) variables was significant with the organ fit decreasing slightly and PMI fit improving. Six genera were determined to be driving the ordination by environmental fit (**Figure 5B**).

Differential Expression

The edgeR mean transcript library sizes after filtering for the intestines was 24,055 (min = 10,549, max = 38,257, SD = 8702.23), the heart was 116,150 (min = 52,102, max = 191,191, SD = 47,094.28), and the bone marrow was 22,882 (min = 17,500, max = 84,116, SD = 25,909.62). The DESeq2 mean transcript library sizes after filtering for the intestines was 17,899 (min = 9570, max = 26,012, SD = 4920.05), the heart was 110,393 (min = 50,531, max = 185,198, SD = 46,014.13), and the bone marrow was 34,932 (min = 14,795, max = 79,693, SD = 24,868.38). Significantly differentially expressed transcripts for each method were determined, but only transcripts that were reported as significant by both edgeR and DESeq2 were considered significant to reduce program bias (**Table 2**). The intestines contained no significantly down-regulated or up-regulated transcripts.

In total, the heart contained 58 significantly down-regulated transcripts, and 8305 significant up-regulated transcripts.

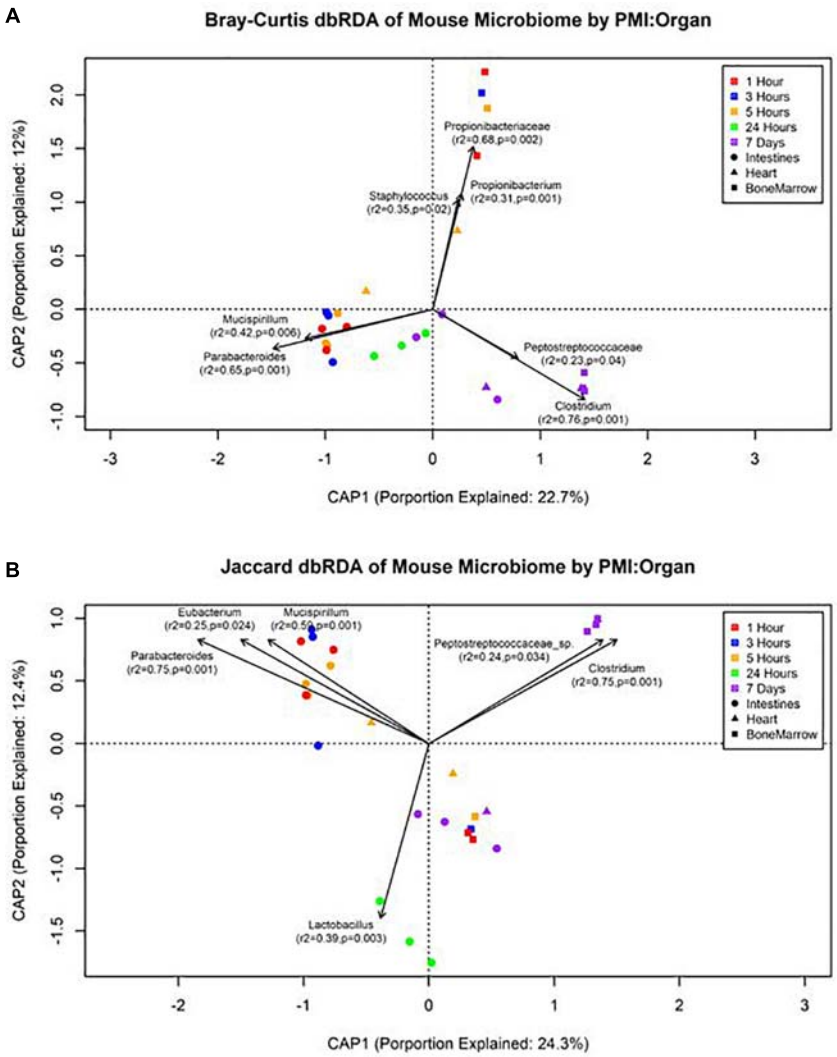


FIGURE 5 | Distance-based RDA plots of the mouse organ microbiome over the postmortem interval. Distance-based RDA plots were created for both **(A)** Bray–Curtis and **(B)** Jaccard distances. Colors represent the postmortem intervals and shapes represent the organ from where the sample was obtained. Linear vectors were determined by genera with a significant environmental fit to the plot based on a $p < 0.05$.

Out of the significant transcripts in the heart, 25 of the 58 down-regulated transcripts and 753 of the 8305 up-regulated transcripts were annotated as hypothetical or ribosomal proteins. In total, the bone marrow contained 734 significantly down-regulated transcripts and 22 significant up-regulated transcripts. Out of the significant transcripts in the bone marrow, 422 of the 734 down-regulated transcripts and 12 of the 22 up-regulated transcripts were annotated as hypothetical or ribosomal proteins. A difference was detected in the heart and bone marrow between the early vs. late and middle vs. late group comparisons, while no difference occurred between the early and middle timepoint groups. In the heart, the pathway regulations were nearly identical with the majority of the pathways detected being up-regulated ($N = 7552$) (**Figure 6A**). Metabolic pathways were the most up-regulated with other notable pathways detected in abundance being stress response, sporulation, cell motility, and membrane transport. Few transcripts were significantly down-regulated ($N = 33$), but the most abundantly down-regulated were associated with metabolism ($N = 15$).

In the bone marrow, the majority of the significant transcripts occurred from a comparison of the early and late timepoint groups and most pathways were down-regulated ($N = 312$) (**Figure 6B**). Energy and carbohydrate metabolism were the most detected down-regulated pathways with other notable pathways being transport and catabolism, stress response, and cell motility. Few transcripts were significantly up-regulated ($N = 10$), though the most abundant pathway was sporulation ($N = 3$). A complete list of the individual significant transcripts with their logFC, FDR, NCBI nr protein annotations, and pathway annotations can be found in Online **Supplementary Material** (**Supplementary Table S1**).

TABLE 2 | Number of transcripts detected from differential expression analysis.

Organ	Time comparison	Number of significantly up-regulated transcripts		Number of non-significantly expressed transcripts		Number of significantly down-regulated transcripts	
		DESeq2	EdgeR	DESeq2	EdgeR	DESeq2	EdgeR
Intestines	Early vs. middle	0(0)	0(0)	26,563	26,500	0(0)	2(0)
	Middle vs. late	0(0)	0(0)	26,563	26,501	0(0)	1(0)
	Early vs. late	0(0)	439(0)	26,563	26,055	0(0)	8(0)
Heart	Early vs. middle	86(0)	1(0)	27,968	24,623	52(0)	1(0)
	Middle vs. late	6024(3774)	6358(3774)	27,968	17,740	41(14)	527(14)
	Early vs. late	6023(3778)	8919(3778)	27,968	13,613	44(19)	2093(19)
Bone marrow	Early vs. middle	1(0)	0(0)	14,329	10,416	687(0)	0(0)
	Middle vs. late	36(3)	17(3)	10,278	14,326	483(44)	121(44)
	Early vs. late	16(7)	1022(7)	8632	14,689	704(268)	762(268)

Time comparisons are separated by their postmortem time groups: early (1 h, 3 h, 5 h), middle (24 h), and late (7 d). The number of significant transcripts from each analysis method is shown with the number of common transcripts between the two methods being in parenthesis.

DISCUSSION

The microbial changes that take place during decomposition have been previously studied with the intent to describe patterns that are consistent, reproducible, and precise for forensic evidence. Our data have shown that an individual *S. aureus* bacterial strain can be tracked as it migrates across organs in the body and behaves in a similar manner across different locations for when it reached maximum abundance and subsequently began to decline. *S. aureus* KUB7 and *C. perfringens* WAL 14572 introduction nor surface sterilization significantly altered the bacterial community structures. It is possible the introduction of *S. aureus* KUB7, while detectable through sensitive DNA techniques, was not at a high enough concentration to highly alter the existing community structure in a manner that would affect the diversity analyses. *C. perfringens*, while not detectable with qPCR, did not appear at levels higher than rare taxa during early decomposition in organs in which *C. perfringens* is not considered natural microbiota, such as the heart. A limitation to this approach of colonization is the difficulty to accurately assess the extent of competition that may have occurred between the natural microbiota and introduced species, and although the introduction of new species to any environment will inevitably affect the ecosystem, our results showed minimal microbial structure disruption when comparing the inoculated versus uninoculated mice suggesting that future transmigration studies could be performed without major concern for causing detrimental effects to the natural microbiota as long as researchers use appropriate organisms and colonization levels.

Additionally, the lack of significant microbial community alterations associated with internal organs by surface sterilization suggests that while external microbiota play a large role in skin decomposition, the breakdown of internal tissues relies predominately on internal microbes. For example, if after sterilization of the skin, internal organs have similar microbial profiles to non-sterilized hosts then the skin microbiome is not playing a large role in the internal organ decomposition. Alternatively, if skin-sterilized mice had drastically different microbial profiles internally than their non-sterilized counterparts, it would suggest that the microbial communities found in the internal organs during decomposition are heavily affected by the transmigration of skin microbes to the internal organs. Our results showed the former, suggesting that internal organs are decomposed primarily by internal microbes and not skin microbes. While this is not surprising, few studies to our knowledge have specifically tested this hypothesis. These results provide strength for the use of internal microbes as potential internal PMI biomarkers, and are useful for future studies that may aim to investigate transmigration and source tracking of specific internal microbes in relation to PMI.

Our transmigration tracking of inhaled *S. aureus* KUB7 showed similar overarching trends in the body as a whole, but with slight differences depending upon the organ (**Figure 3**). In the lungs (**Figure 3A**) and the heart (**Figure 3C**), *S. aureus* KUB7 reached its highest count at 14 d, while in the intestines (**Figure 3B**) and the bone marrow (**Figure 3D**) the highest count was reached at 7 d. We expected the lungs and the intestines to contain *S. aureus* KUB7 immediately after death as these sites were most likely to be colonized following inhalation, and sometimes leading to intake of *S. aureus* KUB7 down the esophagus. The decrease of *S. aureus* KUB7 in the lungs immediately following death could be attributed to competition with other microbial species, an increase in host immune gene transcripts which occurs up to 24 h postmortem, and/or lack of oxygen as blood flow and breathing ceased, though further research is needed to discover the specific mechanisms taking place in our system (Coleman et al., 1983; Pozhitkov et al., 2017). *S. aureus* is a facultative anaerobe that has been shown to have much slower growth in the absence of oxygen, though the lack of oxygen does not completely stop growth (Coleman et al., 1983; Belay and Rasooly, 2002). This may account for the decrease immediately following death, as metabolism shifted and a lag of growth occurs. The spike of growth from 7 to 14 d is likely due to the reintroduction of oxygen, being accustomed

to an anaerobic environment, complete immune system loss, or increase of nutrients present after tissue breakdown (Janaway et al., 2009; Hyde et al., 2013; Crippen et al., 2015; Pozhitkov et al., 2017). The same reasoning can be used to explain the trend found in the intestines as the growth spike occurs at the same time, but the absence of a dip immediately following death may be attributed to the fact that the environment in the intestines is already anaerobic and a drastic shift in oxygen availability did not occur coupled with the fact the *S. aureus* KUB7 present had likely already adjusted to this environment during living colonization. Although it is important to note that we did not generate oxygen level data from our model and could not confirm the aerobic and anaerobic shifts. Furthermore, the presence of non-circulating, active but decreasing host apoptotic cells and neutrophils along with increased immunity gene transcripts between early and late timepoints could also potentially account for fluctuations in *S. aureus* KUB7 concentrations (Heimesaat et al., 2012; Pozhitkov et al., 2017). All organs returned to bacterial levels either similar to or below 1 h postmortem levels by the end of 7 d.

The bone marrow and heart trends are interesting as they do not represent locations where colonization should have occurred from initial host inhalation, and in fact should represent sterile locations during life that are colonized through transmigration after death. Interestingly, *S. aureus* KUB7 was detected within the first hour in the bone marrow, possibly resulting from rapid transmigration or an artifact from the skin when the femur was cut. If rapid transmigration occurred, then *S. aureus* KUB7 was not able to thrive in the early bone marrow environment leading to the immediate decrease at 3 h. However, transmigration began slowly around 5 h until reaching its highest concentration after 7 d. The heart had no detection of *S. aureus* KUB7 until 7 d postmortem and reached maximum detected abundance at 14 d. Suggesting *S. aureus* KUB7 migrated to the heart sometime after 24 h, though it is not clear exactly when it occurred. The 14 d peak may be due to the lag in time between transmigration and heart colonization, as the organ is presumed sterile during host life. The decline seen at the later timepoints is likely due to nutrient limitations or toxicity. However, determining the window where migration into the heart can be pinpointed may serve as a useful biological marker for narrowed PMI estimation ranges.

Bacterial community and differential expression analyses aimed to (1) strengthen existing knowledge of community shifts in vertebrate intestines, (2) provide data toward community shifts in underexplored organs, and (3) identify cellular processes and metabolic changes taking place from early to late decomposition. We provide one of the first sets of metatranscriptomic analyses of these communities showing genes active across decomposition time. Functional analyses performed by other groups have been based on 16S rRNA predictive function, but it is important to study active gene function because, while it may be predicted a gene is expressed, it must be confirmed by expression data if functional biomarkers can be discovered. Additionally, the ability to determine functional activity in a microbial community has the opportunity to by-pass one of the main concerns of solely monitoring the postmortem microbial structure. This concern focuses on the fact that human microbial profiles are unique to the individual and while one individual may carry a certain species of bacteria, another person may not. This can be concerning if using detailed species markers that may not be found ubiquitously across human populations. Monitoring the function of microbes during decomposition may circumvent this concern because, while the individual species may change between people, the functions needed to survive during each PMI should be redundant. For example, the functional pathways needed to utilize decomposition nutrients and survive oxygen fluxes will be highly expressed no matter the individual organism. Another consideration is that species in microbial communities may have dissimilar functional roles during decomposition, that might be obscured while analyzing functional potential from metagenomic data. Therefore, integrating community composition with active gene function may provide evidence to the state of the microbial decomposition ecology across PMIs that are more concrete and less biased to the individual. This coupling of structure and function data, as we have shown, may also yield a predictive framework for determining associations between

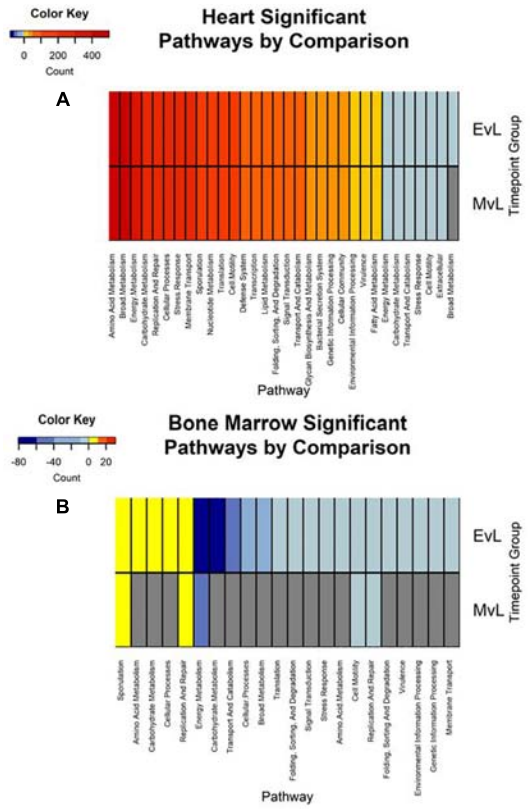


FIGURE 6 | Heatmaps of the significant pathway regulation during timepoint group comparisons. Heatmaps for **(A)** heart and **(B)** bone marrow representing the transcript count of each pathway annotated from the significant transcripts of each comparison by color. Pathway is included on the *x*-axis and timepoint group (EvL = early vs. late, MvL = middle vs. late) on the *y*-axis. Down-regulated transcripts were considered negatives and up-regulated transcripts were considered positives. Pathways that were not detected in a comparison are in gray.

community structure and function during decomposition, with utility for forensic science. Therefore, we suggest that more empirical data should be gathered linking microbial functional groups and genes with community structure in studies of decomposition ecology.

Our data suggest genera richness and Shannon diversity indices vary by organ. Bray–Curtis and Jaccard indices were affected by both organs and postmortem time. This is not surprising as the intestines are a natural location of high levels of microbial diversity compared to other organs, and as decomposition occurs these less diverse organs begin to increase in diversity as transmigrations occur due to the increase of nutrients and lack of immune system after 24 h (Pechal et al., 2014; Guo et al., 2016). These microbial shift trends have been shown in both animal and human models (Heimesaat et al., 2012; Hyde et al., 2013). We found that Bray–Curtis dissimilarity index ordination fit the community profile and genera environmental fit better than Jaccard (**Figure 5**). Jaccard indices treat data as presence–absence, while Bray–Curtis indices accounts for genera richness, which is needed when comparing organs and postmortem times that have been previously shown to vary drastically in their diversity and richness.

We detected distinct microbial changes in the community structure, which corroborate shifts seen in other studies (Metcalf et al., 2013; Tuomisto et al., 2013; Pechal et al., 2014; Hauther et al., 2015). In particular, the ratio decreased of *Parabacteroides* and increase of *Lactobacillus* in the first 24 h followed by the increased ratio of *Enterococcus* and *Clostridium* in late stages is a common postmortem trend (**Figure 4B**). The Bray–Curtis ordination (**Figure 5A**) similarly shows the transition of the intestinal community from *Parabacteroides* to *Clostridium*. However, we did not detect any significantly differential expressed transcripts within the timepoint group comparisons for the intestine microbial communities (**Table 2**). The cellular pathways needed to survive in the intestines may be relatively consistent during the first 7 d, but then community turnover in the intestines may be attributed to factors involved in abilities to use alternate sources to maintain existing pathways, outcompete for nutrients, or replicate faster.

Due to poor sequence coverage or bacterial abundance lower than the needed threshold for quality sequencing, only two lung samples (MG36 and MG60) and two heart samples (MG30 and MG34) from the first 24 h could be classified for structure profiles (**Figures 4A,C**). The two 5 h cardiac samples were drastically different because MG30 only has a single genus detected (*Escherichia*), while MG34 and contained six genera that were not considered rare taxa. The 7 d heart samples were mostly predominated by *Clostridium* with one sample containing *Lactobacillus* and Peptostreptococcaceae; all of which are common Gram-positive gut bacteria associated with decomposition (Pechal et al., 2014; Javan et al., 2017). The heart contained the most differentially expressed transcripts (**Table 2**). Over 99% of the transcripts were up-regulated from ≤24 h to 7 d postmortem. This activity is likely due to the influx of microorganisms from transmigration leading to an increased usage of motility and metabolism pathways to use nutrients in the new environment (**Figure 6A**) as some organisms, such as some species of *Clostridium*, are motile and replicate extremely fast. These organisms could also have migrated to the heart between 24 h and 7 d, leading to nutrient depletion after 7 d with an increased stress response and sporulation to prepare for dormancy.

The bone marrow had bacterial classifications in five samples ≤24 h postmortem (**Figure 4D**). The majority of these early-detected genera are associated with the skin microbiome. Since early timepoints contained multiple genera associated with the skin microbiome, it is likely early bone marrow samples obtained these genera during dissection, but it cannot be fully ruled out that early transmigration took place. As with the heart, the bone marrow 7 d samples were dominated by *Clostridium*. However, opposed to the heart, the bone marrow's differentially expressed transcripts were predominately down-regulated (312 of 322 classified genes) (**Table 2**). These down-regulated transcripts were primary attributed to energy metabolism, carbohydrate metabolism, and transport and catabolism (**Figure 6B**). The early bone marrow genera can survive in oxygenated environments (i.e., *Propionibacterium*) so that anaerobic shifts would cause oxidative energy metabolism pathways to be down-regulated as they switch to the preferred anaerobic energy metabolism pathways. Additionally, these profiles could result from late-arriving genera (i.e., *Clostridium*) having to use carbohydrates not available in the early stages of decomposition, and obtaining energy anaerobically as the oxygen availability shifted from aerobic to anaerobic.

We also detected up-regulation of sporulation in marrow samples during late decomposition, suggesting a need to prepare for dormancy, similar to the heart community. The Bray–Curtis ordination (**Figure 5A**) separated the early bone marrow samples from the rest of the organs due to the early classification of potential skin microbes, but showed similarities in bone marrow and heart microbial communities. A homogenization of microbial communities (*Clostridium* begins to dominate) probably occurs as organ communities become more similar due to the decomposition process. However, a similar trend was not clearly visualized by the Jaccard ordination (**Figure 5B**), as early bone marrow and late intestines were similar in Jaccard distance though not reflected in the detected microbial communities from each organ at those times. These conflicting data reflect an important point of consideration when dealing with different organ ecosystems using a presence–absence model, as opposed to taking in account the genera richness. This suggests comparing ecosystems suspected to vary in microbial abundance and diversity should be compared using methods that take into account richness, to not overinflate rare taxa (Lozupone et al., 2007).

Additionally, it is important to note that the decomposition process is highly affected by environmental factors and host individuality. In this study, many of these factors were controlled by using a genetically identical model, eliminating scavengers, and maintaining ambient temperature/moisture. But in a more natural setting these factors alter decomposition rates

making PMI estimation more difficult. Although the time it takes a cadaver to progress through the stepwise patterns of decomposition can vary, the process itself is rather ubiquitous, with the exception of extreme cases. Because of this, it is suspected that metabolic activities of microorganisms associated with decomposition may be similar across ecosystems and individuals during decomposition with little variance. For example, the need to transcribe genes for survival during oxygen fluxes is universal no matter the specific species present. Our data show that the functional profiles of these microbial communities are PMI dependent and coupled with community structural data may provide better insight on the mechanisms of microbially mediated decomposition, though the impact of environmental factors on these functional changes are still poorly understood.

We have also shown community structure results in our system that correlate with results that have been shown in work by other researchers using animal and human models in multiple geographic locations (Pechal et al., 2014; Guo et al., 2016; Metcalf et al., 2016; Javan et al., 2017). This corroboration by multiple, independent sources is a critical step for future implementation of microbiological assessment to the PMI in forensic cases (Burcham et al., 2016). However, abiotic factors not measured in this study, such as pH and oxygen levels, should be assessed in further studies to expand on the microbial physiological responses of these communities. Lastly, we have performed one of the first studies combining bacterial taxonomic and metatranscriptomic analyses as decomposition occurs, which includes detailed exploration of both commonly and underreported microbial communities of the postmortem microbiome. From these, we have shown distinct changes in microbial community structure and function during decomposition succession. Microbial community structure in conjunction with community functioning are imperative to understand as we explore in-depth analysis of transmigration and microbial succession of commensal organisms in response to environmental disturbances within the once living host. We have provided a broad overview of the metabolic and stress response changes taking place during decomposition, along with the individual transcript fold changes, so that future research can narrow its scope to clusters of genes, along with associated community composition, with potential for biomarkers to aid in PMI determination.

ETHICS STATEMENT

This study was carried out in accordance with the principles of the Basel Declaration and recommendations of the Mississippi State University Institutional Animal Care and Use Committee approved (protocol 14–102), Institutional Biosafety Committee (protocol 009–14), and approved by both the University Institutional Animal Care and Use Committee and Institutional Biosafety Committees.

AUTHOR CONTRIBUTIONS

ZB performed quantitative and computational analysis. JP, CS, MB, and HJ conceived of the presented idea, obtained funding for the project, and supervised the findings of this work. JB created the fluorescent strain utilized in this project. JR was responsible for obtaining the metagenomic and metatranscriptomic sequencing data. All authors discussed the results, reviewed, and contributed to the final manuscript.

FUNDING

Funding was provided by Mississippi State University and the National Institute of Justice (2014-DN-BX-K008). Opinions or points of view expressed represent a consensus of the authors and do not necessarily represent the official position or policies of the United States Department of Justice. All demultiplexed raw sequence reads are available on the NCBI Short Read Archive (SRP158407) through BioProject (PRJNA485410).

ACKNOWLEDGMENTS

We would like to thank Courtney Baugher, Caitlyn Cowick, and Andrew Bryant for their work assisting in nucleic acid extractions and quantitative PCR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00745/full#supplementary-material

REFERENCES

Belay, N., and Rasooly, A. (2002). Staphylococcus aureus growth and enterotoxin a production in an anaerobic environment. *J. Food Prot.* 65, 199–204. doi: 10.4315/0362-028X-65.1.199

Benbow, M., Barton, P., Ulyshen, M., Beasley, J., DeVault, T., Strickland, M., et al. (2018). Necrobiome framework for bridging decomposition ecology of autotrophically and heterotrophically-derived organic matter. *Ecol. Monogr.* 89:e01331. doi: 10.1002/ecm.1331

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

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Burcham, Z. M., Hood, J. A., Pechal, J. L., Krausz, K. L., Bose, J. L., Schmidt, C. J., et al. (2016). Fluorescently labeled bacteria provide insight on post-mortem microbial transmigration. *Forensic Sci. Int.* 264, 63–69. doi: 10.1016/j.forsciint.2016.03.019

Can, I., Javan, G. T., Pozhitkov, A. E., and Noble, P. A. (2014). Distinctive thanatomicrobiome signatures found in the blood and internal organs of humans. *J. Microbiol. Methods* 106, 1–7. doi: 10.1016/j.mimet.2014.07.026

Carter, D. O., Yellowlees, D., and Tibbett, M. (2007). Cadaver decomposition in terrestrial ecosystems. *Naturwissenschaften* 94, 12–24. doi: 10.1007/s00114-006-0159-1

Coleman, G., Garbutt, I. T., and Demnitz, U. (1983). Ability of a Staphylococcus aureus isolate from a chronic osteomyelitic lesion to survive in the absence of air. *Eur. J. Clin. Microbiol.* 2, 595–597. doi: 10.1007/BF02016574

Crippen, T. L., Benbow, M. E., and Pechal, J. L. (2015). "Microbial interactions during carrion decomposition," in *Carrion Ecology, Evolution, And Their Applications*, eds M. E. Benbow, J. K. Tomberlin, and A. M. Tarone (Boca Raton, FL: CRC Press), 31–64. doi: 10.1201/b18819-5

Damann, F., and Carter, D. (2013). "Human decomposition ecology and postmortem microbiology," in *Manual of Forensic Taphonomy*, eds S. A. Symes and J. T. Pokines (Boca Raton, FL: CRC Press).

Delgado-Baquerizo, M., Giaramida, L., Reich, P., Khachane, A., Harmonts, K., Edwards, C., et al. (2016). Lack of functinoal redundancy in the relationship between microbial diversity and ecosystem functioning. *J. Ecol.* 104, 936–946. doi: 10.1111/1365-2745.12585

Finley, S. J., Benbow, M. E., and Javan, G. T. (2015). Microbial communities associated with human decomposition and their potential use as postmortem clocks. *Int. J. Legal Med.* 129, 623–632. doi: 10.1007/s00414-014-1059-0

Fox, J., and Weisbert, S. (2011). *An R Companion to Applied Regression*. Thousand Oaks, CA: SAGE publications.

Guo, J., Fu, X., Liao, H., Hu, Z., Long, L., Yan, W., et al. (2016). Potential use of bacterial community succession for estimating post-mortem interval as revealed by high-throughput sequencing. *Sci. Rep.* 6:24197. doi: 10.1038/srep24197

Hauther, K. A., Cobaugh, K. L., Jantz, L. M., Sparer, T. E., and DeBruyn, J. M. (2015). Estimating time since death from postmortem human gut microbial communities. *J. Forensic Sci.* 60, 1234–1240. doi: 10.1111/1556-4029.12828

Heimesaat, M. M., Boelke, S., Fischer, A., Haag, L. M., Loddenkemper, C., Kuhl, A. A., et al. (2012). Comprehensive postmortem analyses of intestinal microbiota changes and bacterial translocation in human flora associated mice. *PLoS One* 7:e40758. doi: 10.1371/journal.pone.0040758

Herve, M. (2018). *RVAideMemoire: Testing and Plotting Procedures for Biostatistics. R package Version 0.9-69-3 ed*. Available at: https://cran.r-project.org/web/packages/RVAideMemoire/RVAideMemoire.pdf (accessed March 26, 2019).

Hyde, E., Metcalf, J., Bucheli, S., Lynne, A., and Knight, R. (2017). "Microbial communities associated with decomposing corpses," in *Forensic Microbiology*, ed. D. O. Carter (Hoboken, NJ: Wiley). doi: 10.1002/9781119062585.ch10

Hyde, E. R., Haarmann, D. P., Lynne, A. M., Bucheli, S. R., and Petrosino, J. F. (2013). The living dead: bacterial community structure of a cadaver at the onset and end of the bloat stage of decomposition. *PLoS One* 8:e77733. doi: 10.1371/journal.pone.0077733

Janaway, R. C., Percival, S. L., and Wilson, A. S. (2009). "Decomposition of Human Remains," in *Microbiology and Aging*, ed. S. L. Percival (Berlin: Springer), 313–334. doi: 10.1007/978-1-59745-327-1_14

Javan, G. T., Finley, S. J., Can, I., Wilkinson, J. E., Hanson, J. D., and Tarone, A. M. (2016). Human thanatomicrobiome succession and time since death. *Sci. Rep.* 6:29598. doi: 10.1038/srep29598

Javan, G. T., Finley, S. J., Smith, T., Miller, J., and Wilkinson, J. E. (2017). Cadaver thanatomicrobiome signatures: the ubiquitous nature of clostridium species in human decomposition. *Front. Microbiol.* 8:2096. doi: 10.3389/fmicb.2017.02096

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., et al. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* 46, D335–D342. doi: 10.1093/nar/gkx1038

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kellerman, G. D., Waterman, N. G., and Scharefenberger, L. F. (1976). Demonstration in vitro of postmortem bacterial transmigration. *Am. J. Clin. Pathol.* 66, 911–915. doi: 10.1093/ajcp/66.5.911

Kopylova, E., Noe, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217. doi: 10.1093/bioinformatics/bts611

Krueger, F. (2015). *TrimGalore: A Wrapper Around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to Fast Q Files*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. (accessed January 12, 2017).

Langenheder, S., Lindstrom, E. S., and Tranvik, L. J. (2006). Structure and function of bacterial communities emerging from different sources under identical conditions. *Appl. Environ. Microbiol.* 72, 212–220. doi: 10.1128/AEM.72.1.212-220.2006

Legendre, P., and Anderson, M. J. (1999). Distance-based redundance analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* 69, 1–24. doi: 10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2

Louca, S., Polz, M. F., Mazel, F., Albright, M. B. N., Huber, J. A., O'Connor, M. I., et al. (2018). Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* 2, 936–943. doi: 10.1038/s41559-018-0519-1

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* 73, 1576–1585. doi: 10.1128/AEM.01996-06

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

Melvin, J. R. Jr., Cronholm, L. S., Simson, L. R. Jr., and Isaacs, A. M. (1984). Bacterial transmigration as an indicator of time of death. *J. Forensic Sci.* 29, 412–417. doi: 10.1520/JFS11687J

Metcalf, J. L., Wegener Parfrey, L., Gonzalez, A., Lauber, C. L., Knights, D., Ackermann, G., et al. (2013). A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *eLife* 2:e01104. doi: 10.7554/eLife.01104

Metcalf, J. L., Xu, Z. Z., Weiss, S., Lax, S., Van Treuren, W., Hyde, E. R., et al. (2016). Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science* 351, 158–162. doi: 10.1126/science.aad2646

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2018). *vegan: Community Ecology Package. R Package Version 2.5-2 ed*. Available at: https://cran.r-project.org/web/packages/vegan/vegan.pdf (accessed February 4, 2019).

Pechal, J. L., Crippen, T. L., Benbow, M. E., Tarone, A. M., Dowd, S., and Tomberlin, J. K. (2014). The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. *Int. J. Legal Med.* 128, 193–205. doi: 10.1007/s00414-013-0872-1

Pechal, J. L., Schmidt, C. J., Jordan, H. R., and Benbow, M. E. (2018). A large-scale survey of the postmortem human microbiome, and its potential to provide insight into the living health condition. *Sci. Rep.* 8:5724. doi: 10.1038/s41598-018-23989-w

Pozhitkov, A. E., Neme, R., Domazet-Loso, T., Leroux, B. G., Soni, S., Tautz, D., et al. (2017). Tracing the dynamics of gene transcripts after organismal death. *Open Biol.* 7:160267. doi: 10.1098/rsob.160267

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

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi: 10.1093/nar/gkw1099

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, 902–903. doi: 10.1038/nmeth.3589

Tuomisto, S., Karhunen, P. J., Vuento, R., Aittoniemi, J., and Pessi, T. (2013). Evaluation of postmortem bacterial migration using culturing and real-time quantitative PCR. *J. Forensic Sci.* 58, 910–916. doi: 10.1111/1556-4029.12124

Westreich, S. T., Treiber, M. L., Mills, D. A., Korf, I., and Lemay, D. G. (2018). SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinformatics* 19:175. doi: 10.1186/s12859-018-2189-z

Williamson, H. R., Mosi, L., Donnell, R., Aqqad, M., Merritt, R. W., and Small, P. L. (2014). Mycobacterium ulcerans fails to infect through skin abrasions in a guinea pig infection model: implications for transmission. *PLoS Negl. Trop. Dis.* 8:e2770. doi: 10.1371/journal.pntd.0002770

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 42, D643–D648. doi: 10.1093/nar/gkt1209

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593

Conflict of Interest Statement: 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.

Copyright © 2019 Burcham, Pechal, Schmidt, Bose, Rosch, Benbow and Jordan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

REVIEW
published: 18 April 2019
doi: 10.3389/fmicb.2019.00827

Application of Machine Learning in Microbiology

Kaiyang Qu[1], Fei Guo[1], Xiangrong Liu[2], Yuan Lin[2,3] and Quan Zou[4,5]**

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China, [2] School of Information Science and Technology, Xiamen University, Xiamen, China, [3] Department of System Integration, Sparebanken Vest, Bergen, Norway, [4] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [5] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China

Microorganisms are ubiquitous and closely related to people's daily lives. Since they were first discovered in the 19th century, researchers have shown great interest in microorganisms. People studied microorganisms through cultivation, but this method is expensive and time consuming. However, the cultivation method cannot keep a pace with the development of high-throughput sequencing technology. To deal with this problem, machine learning (ML) methods have been widely applied to the field of microbiology. Literature reviews have shown that ML can be used in many aspects of microbiology research, especially classification problems, and for exploring the interaction between microorganisms and the surrounding environment. In this study, we summarize the application of ML in microbiology.

Keywords: microorganisms, classification, environment, species, association, diseases

OPEN ACCESS

Edited by:
Hongsheng Liu,
Liaoning University, China

Reviewed by:
Yen-Wei Chu,
National Chung Hsing University, Taiwan
Mohamed Elhoseny,
Mansoura University, Egypt

****Correspondence:***
Yuan Lin
linyuan1979@gmail.com
Quan Zou
zouquan@nclab.net

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

Received: *31 January 2019*
Accepted: *01 April 2019*
Published: *18 April 2019*

Citation:
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

INTRODUCTION

Microorganisms first appeared approximately 3.5 billion years ago, making them one of the earliest living things on Earth (Nannipieri et al., 2010). Microorganisms include bacteria, viruses, fungi, some small protozoa, and microscopic algae. These organisms, which are closely related to human beings (Ley et al., 2006a), have a wide range of beneficial and harmful uses, including in the food (Cotter et al., 2005), medicine (Petrof et al., 2012; Yu et al., 2018), agriculture (Morris et al., 1986), industrial (Souza, 2010), environmental protection and other fields (Reiff and Kelly, 2010).

Microbiology is a discipline that studies the structure and function of microbial groups, the interrelationships and mechanisms of internal communities, and the relationships between microorganisms and their environments or hosts (Alexander, 1962; Niel, 1966). The microbiome is a collection of all microbial species and their genetic information and functions in a given environment. Studies of the microbiome also include the interaction between different microorganisms (DiMucci et al., 2018), the interaction between microorganisms and other species (Xie et al., 2018), and the interaction between microorganisms and the environment (Moitinho-Silva et al., 2017). Because of their small size, the microscope is an important tool for studying microorganisms. However, microscopy analyses only allow observation and must therefore be complemented by culture techniques to study the biological, physiological, genetic, metabolic, pathogenic and other biological characteristics of microorganisms (Waldron, 2018). During cultivation, researchers can also explore the interactions between microorganisms and the environment, which reflect the breadth and diversity of microbial distribution. A variety of microorganisms living in different environments or in different hosts form microbial communities, which have extensive and complex interactions with the environment and the host and form various types of ecosystems (Srinivasan et al., 2012; Xie et al., 2018).

With the development of microbial sequencing in recent years, the microbiome has become increasingly popular in many studies. High-throughput sequencing technology has resulted in

generation of an increasing amount of microbial data. Traditional methods using microscopes and biological cultures are expensive and labor intensive; therefore, machine-learning methods have been gradually applied to microbial studies (Huang Y. A. et al., 2017; Huang Z. A. et al., 2017; Wang et al., 2017; Wei et al., 2017a,b; Peng et al., 2018; Yang et al., 2018b; Zou et al., 2018a). Here, we introduce the application of machine learning (ML) in microbial analyses. Since ML is mainly applied to classification and interaction problems, we focus on these two areas. **Figure 1** shows the framework of this paper.

MACHINE LEARNING METHODS

Machine Learning is a multi-disciplinary subject involving many disciplines including probability theory, statistics, approximation theory, convex analysis, and algorithm complexity theory (Qu et al., 2017; Zou et al., 2018b). ML methods can be divided into two types (Zitnik et al., 2019), supervised learning and unsupervised learning. Supervised learning (Stoter et al., 2019) requires that the model be trained using a training set. The training sets for supervised learning include features and results. Common supervised learning algorithms include regression analysis and statistical classification. Unsupervised learning, also known as clustering, adopts k-means to establish a centriole and reduce error through iteration and descent to achieve classification. With the development of ML, more and more fields have begun to use this technique for research (Chen W. et al., 2016, Chen et al., 2017a,d, 2018a,b,e,f,g; Li et al., 2016; Zou et al., 2016, 2017; Ding et al., 2017a,b; Feng et al., 2017a; Yu et al., 2017a; Zeng et al., 2017a, 2018; Liu et al., 2018; Pan et al., 2018; Wei et al., 2018a,b; Yang et al., 2018a; Zhao et al., 2018b; He et al., 2019; Zhang et al., 2019), for example, drug repositioning (Yu et al., 2016b, 2017b), disease-related microRNA (Chen and Huang, 2017; Chen et al., 2017d, 2018b,e,g; Zhao et al., 2018a,c) identification, and disease-related long non-coding RNA identification (Chen and Yan, 2013; Chen et al., 2017e, 2018c; Hu et al., 2017, 2018). There are four main steps in developing ML algorithms (Oudah and Henschel, 2018). The first step is extraction of the features, which is critical to the ML method (Liu et al., 2015). Then, the operational classification units (OTU) table can be obtained by clustering. Next, important features that can improve the accuracy and efficiency are selected. Finally, a training dataset is used to train the model, after which a test set is used to evaluate the model. The process is summarized in **Figure 2**.

In microbial studies, according to the collected samples, obtaining relevant OTU is an important step in the study of microbial data. OTU is a type of similar microorganisms, which are cluster according to the similarity DNA sequences (Blaxter et al., 2005). In recent years, OTUs are always used for microbial diversity, especially when analyzing small subunit 16S or 18S rRNA datasets (Schmidt et al., 2014). Sequences can be clustered according to their similarity to one another, and the researcher sets the similarity threshold. After OTU clustering and species classification annotation for OTU, the OTU table can be obtained, which contains the OTU types and quantities for each sample, as well as species annotation information for each OTU.

As we know, some microbes have higher data dimensions, so feature dimensionality reduction is also an important part of data processing. There are some common methods for reducing the dimensionality and many studies are about how to reduce the dimensionality. For example, the principal components analysis (PCA) is a common reduction dimensionality method, which is mainly to decompose the covariance matrix to obtain the principal components and their weights (Jolliffe, 2002). PCA is often used to reduce the dimensionality of dataset while maintaining the feature that maximizes the contribution of the variance in the data set. Principal co-ordinates analysis (PCoA) is another common method. After sorting the feature values and the feature vectors, PCoA selects the features, which are in the top digits and the most significant coordinates in the distance matrix can be found (Podani and Miklós, 2002). The result is a rotation of the data matrix. It does not change the mutual positional relationship between the sample points, but only changes the coordinate system.

In microbial studies, supervised learning is always used, especially the support vector machine (SVM) (Feng et al., 2013a, 2017b; Chen X. X. et al., 2016; Yang et al., 2016), and the Naïve Bayes (NB) (Feng et al., 2013b,c), random forest (RF) (Chen et al., 2018d), and *k* nearest neighbor (KNN) methods (Chen et al., 2017c).

The SVM is a generalized linear classifier that can perform binary classification of data employing a decision basis, according to the maximum-margin hyperplane of the learning sample. The SVM can classify non-linear data by the kernel methods (Drucker et al., 2002). SVM is widely used in bioinformatics, such as the prediction of proteins (Xu et al., 2018a,b,c). The NB method (Meena and Chandran, 2009), which is a classification based on Bayes' theory and the independent assumption of features that originate from classical mathematical theory (Rodríguez and Kuncheva, 2007), has a solid mathematical foundation and stable classification efficiency. The NB classifier, which requires only a few parameters, is less sensitive to missing data and simpler than other methods (Jordan, 2008). The RF is a classifier that contains multiple decision trees and its output accords to the voting on each decision tree (Svetnik et al., 2003). KNN (Cui et al., 2001) is a theoretically mature method. The method infers the sample category based on its neighbors. The main steps of the algorithm are as follows (Liao and Vemuri, 2002). First, the distance, which is between the test sample and each training sample, should be calculated. Then, the nearest *k* training samples are found as the nearest neighbors of the test sample. Finally, the test sample is classified according to the categories of the *k* nearest neighbors.

CLASSIFICATION AND PREDICTION IN MICROBIOLOGY

Prediction of Microbial Species

There are two main types of microorganisms (Maiden et al., 1998), one of them with non-cellular morphology (Yeom and Javidi, 2006), such as viruses, and the other with cellular

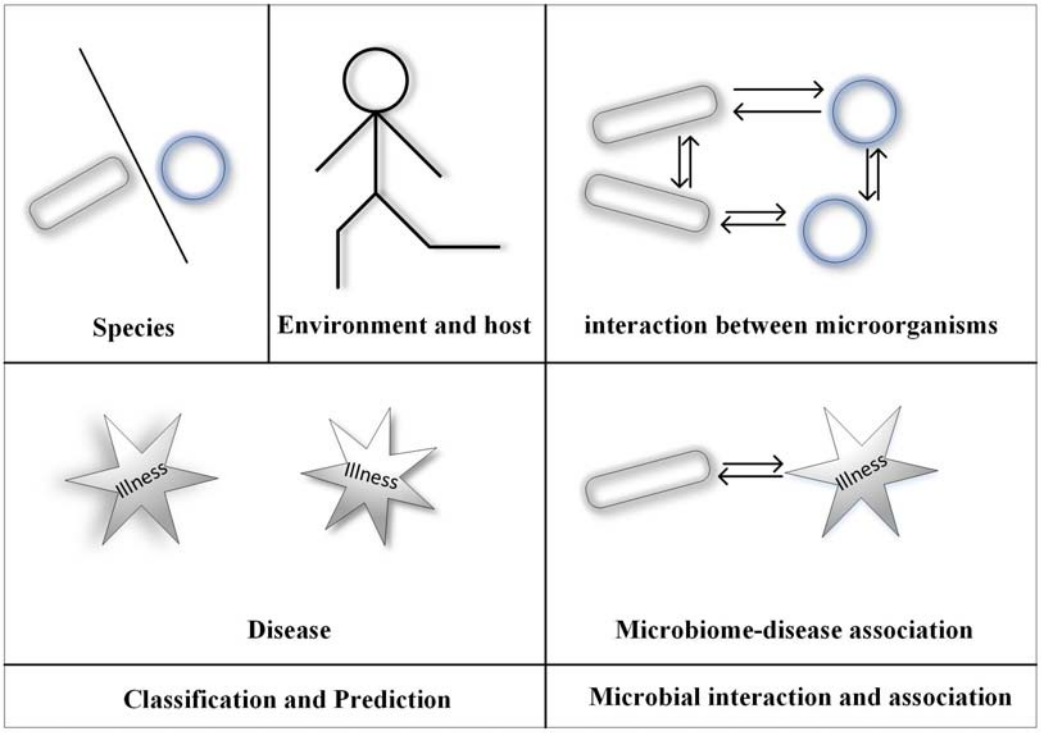


FIGURE 1 | The framework of this paper.

morphology that can divided into two types, one of them namely prokaryotes (Weinbauer, 2010), such as archaea and eubacteria, and the other namely eukaryotes (Nowrousian, 2010), such as fungi and unicellular algae. Different microorganisms have different characteristics, so it is important to identify the microorganisms properly. There are two main approaches to the identification of microorganisms. In one, the species of an unknown microorganism is determined with the goal of classifying it based on its domain, kingdom, phylum, class, order, family genus and species. In the other, the goal is to determine whether an unknown microorganism belongs to a specific species or not. For example, we can determine if an unknown microorganism is a virus or not, or more specifically, whether it is a certain virus. In this section, we will introduce recent studies that have used machine-learning methods to predict microorganisms.

In the study (Murali et al., 2018), the authors classified specific species of microorganisms using the IDTAXA, which employed the LearnTaxa and IdTaxa functions. Both of these functions are part of the R package DECIPHER, which was released under the GPLv3 license as part of the Bioconductor, which provides tools for the analysis and comprehension of high-throughput genomic data. The LearnTaxa function attempts to reclassify each training sequence into its tagged taxon using a method known as tree descent, which is similar to the decision tree, a commonly ML algorithms. IdTaxa uses the objects returned by the LearnTaxa and query sequences as input data. This system returns the classification results for each sequence in the taxonomic form and provides the relevant confidence for each level. If the confidence does not reaches the required value, which indicates that the classification cannot be accurately performed at that level. The classification of IdTaxa may lead to different conclusions in microbiological studies. Although the misclassification is small, many of the remaining misclassifications may be caused by the errors in the reference taxonomy. Fiannaca et al. (2018) presented a method for identifying the 16S short-read sequences based on *k*-mer and deep learning. According to their results, the method can classify both 16S shotgun (SG) and amplicon (AMP) data very well.

It is important to identify specific microbial sequences in mixed metagenomics samples. At present, gene-based similarity methods are popularly used to classify prokaryotic and host organisms from mixed samples; however, these techniques have major weakness. Therefore, many studies have been conducted to identify better methods for identification of specific microorganisms. Amgarten et al. (2018) proposed a tool known as MARVEL for predicting double-stranded DNA bacteriophage sequences in metagenomics. MARVEL uses the RF method, with a training dataset composed of 1,247 phage and 1,029 bacterial genomes and a test dataset composed of 335 bacteria and 177 phage genomes. The authors proposed six features to identify the phages, then used random forests to select features and found three features provided more information (Grazziotin et al., 2017). Ren et al. (2017) developed VirFinder, which is a ML method based on *k*-mer for virus overlap group identification that avoids gene-based similarity searches. VirFinder trains the

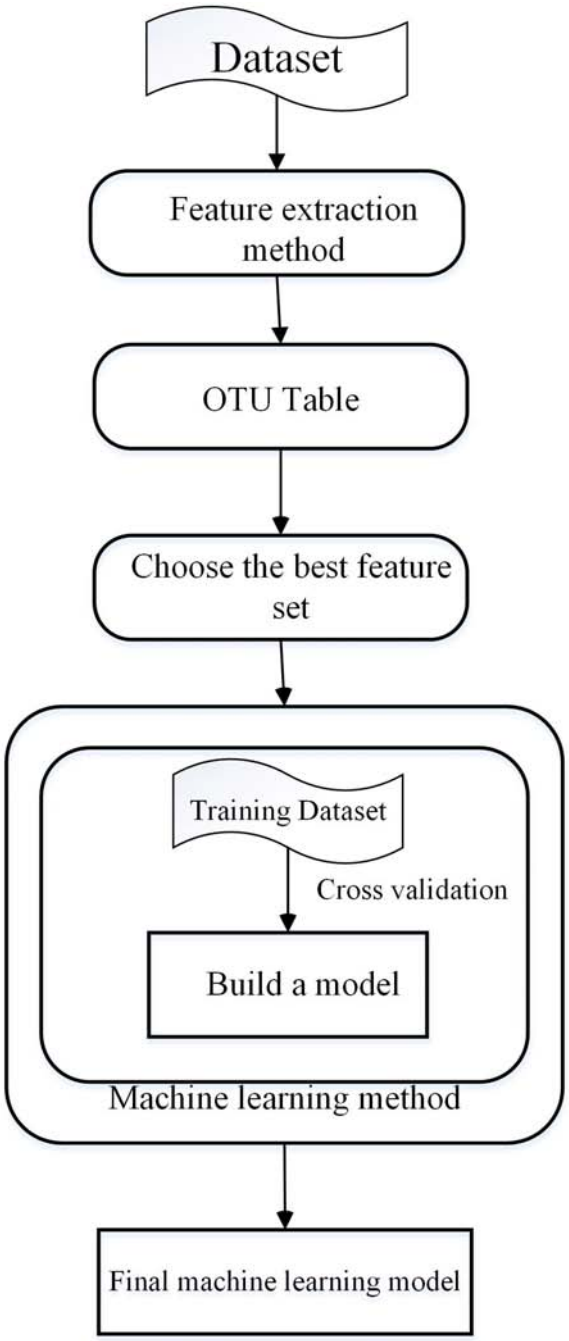


FIGURE 2 | The main steps of machine learning in microbiology.

ML model through known viral and non-viral (prokaryotic host) sequences to detect the specificity of viral *k*-mer frequencies. The model was trained with host and viral genomes prior to January 1, 2014, and the test set consisted of sequences obtained after January 1, 2014. VirSorter (Roux et al., 2015) is based on reference dependence and reference independence in different kinds of microbial sequence data to identify the viral signal. Experimental results have shown that VirSorter has good performance, especially for predicting viral sequences outside the host genome.

TABLE 1 | The available data and materials for prediction of microbial species.

Studies	Availability of data and materials	Reference
IDTAXA	http://DECIPHER.codes	Murali et al., 2018
Fiannaca et al.	https://github.com/IcarPA-TBlab/MetagenomicDC	Fiannaca et al., 2018
MARVEL	https://github.com/LaboratorioBioinformatica/MARVEL	Amgarten et al., 2018
VirFinder	https://github.com/jessieren/VirFinder	Ren et al., 2017
VirSorter	https://github.com/simroux/VirSorter	Roux et al., 2015

The above methods specifically classify microorganisms according to different needs. When we want to know the taxonomy information of microorganisms, we can use the method, which proposed by Murali et al. (2018). Moreover, MARVEL, VirSort, and VirFinder can identify specific types of microorganisms. According to the Amgarten et al. (2018), these three methods have comparable performance on specificity, but MARVEL has a better recall (sensitivity) performance. We have compiled materials for implementation of the above methods, which are shown in **Table 1**.

Prediction of Environmental and Host Phenotypes

With the development of next-generation DNA and high-throughput sequencing, a new area of microbiology has been generated. The main research in this field is to link microbial populations to phenotypes and ecological environments, which can provide favorable support for disease outbreaks and precision medicine (Atlas and Bartha, 1981). It is well known that some microorganisms are parasitic and that the surrounding environment and host cells have an important impact on the microbial population. Differences in nutrient availability and environmental conditions lead to differences in microbial communities (Moran, 2015). Because microorganisms can exchange information with the surrounding environment and host cells, we can predict the environmental and host phenotypes based on the microorganisms that are present (Xie et al., 2018). This provides a more comprehensive understanding of the environment and the host, so that we can better use the environment and protect the host. Many studies have recently been conducted to predict environmental and host phenotypes using microorganisms. In this section, we introduce these studies.

Asgari et al. (2018) used shallow subsample representation based on *k*-mer and deep learning, random forests, and SVMs to predict environmental and host phenotypes from 16S rRNA gene sequencing using the MicroPheno system. They found that the shallow subsample representation based on *k*-mer is superior to OTU in terms of body location recognition and Crohn's disease prediction. In addition, the deep learning method is better than the RF and SVM for large datasets. This method not only can improve the performance, but also avoid overfitting. Moreover, it can reduce the time of pretreatment. Statnikov et al. (2013) used OTUs as an input feature and processed

the data as follows. First, the authors sequenced the original DNA, after which they removed the human DNA sequence and defined the OTUs based on the microbial sequence. Next, they quantified the relative abundance of all sequences belonging to each OTU. The authors used SVM, kernel ridge regression, regularized logistic regression, Bayesian logistic regression, the KNN method, the RF method and probabilistic neural networks with different parameters and kernel functions. Overall, they investigated 18 ML methods. In addition, they used five feature extraction methods. The experimental results revealed that the RF, SVM, kernel-regression and Bayesian logic use Laplacian prior regression provided better performance. Based on their research, human skin microorganisms collected from objects that have been touched can be used to identify the individual from which they originated. In this work, the author used a variety of classification and dimensionality reduction methods to explore the effects of each method. It is very useful for the next work, which provides a comprehensive comparison. Schmedes et al. (2018) used the microbial community for forensic identification. In their study, they developed the hidSkinPlex, a novel targeted sequencing method using skin microbiome markers developed for human identification. In forensic science, it is important to estimate the time of death. Johnson et al. (2016) used KNN regression to predict the time interval after death using datasets from nose and ear samples. This indicates that skin microbiota can be an important tool in forensic death investigation. Traditionally, marine biological monitoring involves the classification and morphological identification of large benthic invertebrates, which requires a great deal of time and money. Cordier et al. (2017) used eDNA metabarcoding and supervised ML to build a powerful prediction model of benthic monitoring. Moitinho-Silva et al. (2017), studied the microbial flora of sponges and their HMA-LMA status demonstrated the applicability of ML to exploring host-related microbial community patterns.

Due to the specificity of microbial communities, we can better identify the environment and the host. Moreover, we can judge the existing environmental conditions and host survival status according to the existence of microbial community. We summarize the available datasets and methods, which are shown in **Table 2**.

Using Microbial Communities to Predict Disease

Microbiomes are important to human health and disease (Bourne et al., 2009). Indeed, there are many microbial communities in the human body. Once a microbial community is out of balance or foreign microorganisms invade, the human body is likely to get sick. For example, intestinal microbial communities are associated with obesity (Ley et al., 2006b) and pulmonary communities with pulmonary infection (Sibley et al., 2008). Because of the complexity of these communities, it is difficult to determine which kind of microbiome communities cause of the disease. Recently, many studies have investigated use of microbiome communities to predict diseases, especially bacterial vaginosis (Srinivasan et al., 2012; Deng et al., 2018) and inflammatory bowel disease (Gillevet et al., 2010). By analyzing microbial communities, we can better understand the disease and then make effective decisions regarding treatment. Therefore, in this section, we discuss current studies investigating use of microbiome communities to predict diseases.

Bacterial vaginosis (BV) is a disease associated with the vaginal microbiome. Beck and Foster (2014) used the genetic algorithm (GP), RF, and logistic regression (LR) to classify BV according to microbial communities. There are two criteria for BV, the Amsel standard, which accord to the discharge, whiff, clue cells, and pH (Amsel et al., 1983), and Nugent score, which dependents on counting gram-positive cells (Nugent et al., 1991). The dataset in Beck et al. study was from Ravel et al. (2011) and Sujatha et al. (2012). The method in the paper (Beck and Foster, 2014) first classifies BV according to vaginal microbiota and related environmental factors, then identifies the most important microbial community for predicting BV.

Hierarchical feature extraction is based on the classification of microbes from kingdoms to species. The existing stratification feature selection algorithm will lead to information loss, and the stratification information of some 16S rRNA sequences is usually incomplete, influencing the classification. Therefore, Oudah and Henschel (2018) proposed a method known as hierarchical feature engineering (HFE) to identify colorectal cancer (CRC). To accomplish this, they used RF, decision trees and the NB method to classify a dataset of Next Generation Sequencing based 16S rRNA sequences provided by metagenomics studies. This method is good for processing datasets with high dimensional features. Moreover, the available dataset and method are in https://github.com/HenschelLab/HierarchicalFeatureEngineering.

In another study (Wisittipanit, 2012), the author focused on predicting inflammatory bowel disease. In that study, patients with Crohn's disease and ulcerative colitis were compared with healthy controls to identify differences between the mucosa and lumen in different intestinal locations. The author used the Relief algorithm (Kira and Rendell, 1992) to select features, and Metastats (White et al., 2009) to detect differential features. Finally, the author used KNN and SVM as classifiers to perform disease specificity and site specificity analysis.

In this section, we discuss using microorganisms to predict different diseases. Beck and Foster (2014) predicted BV according to the microorganisms and the diagnosis standard of BV. HFE identified the CRC according to the OTU ID and the taxonomy information. Wisittipanit proposed a method to predict Crohn's disease, based on OTU and feature selection method. The above methods used different ideas to predict diseases by using microorganisms and obtained good results. This indicates that some diseases affect human colonies. According to these colony changes, we can not only predict the disease, but also treat the

TABLE 2 | The available data and materials for prediction of environmental and host phenotypes.

Studies	Availability of data and materials	Reference
Asgari et al.	https://llp.berkeley.edu/micropheno	Asgari et al., 2018
Statnikov et al.	https://link.springer.com/article/10.1186/2049-2618-1-11	Statnikov et al., 2013

disease according to the colony condition, which is a direction for future research.

INTERACTION AND ASSOCIATION IN MICROBIOLOGY

Interaction Between Microorganisms

The collective behavior of microbial ecosystems in biomes is the result of many interactions between community members. These interactions include metabolite exchange, signaling and quorum sensing processes, as well as growth inhibition and killing (Langille et al., 2013; DiMucci et al., 2018). Understanding the interspecific interactions within microbial communities is critical to understanding the functions of natural ecosystems and the design of synthetic consortia (Mainali et al., 2017). Therefore, in this section, we introduce the application of ML to investigation of interactions between microorganisms.

DiMucci et al. (2018) showed how the microbial interaction network can be combined with the characteristic level of individual microbes to provide an accurate inference of the missing edges in the network and a constructive mechanism of the interaction. The same authors proposed the notion of a composite vector that combined the generated trait vectors and pairwise interactions. The training set for the model is all observed interactions. The model was then used to predict the unobserved interactions. If the random forest classifier is used, feature contributions can be calculated. Microbial interactions in the soil can affect crop yields; therefore, Chang et al. (2017) used the random forest method to predict the productivity based on the microorganisms. In this study, the improved crop productivity differences were linked to the soil microbial composition.

There are cooperative and competitive relationships within the same microbial population. Moreover, there are eight relationships between the different microbial populations, which are neutralism, commensalism, synergism, mutualism, competition, amensalism, parasitism and predation. Understanding the interactions between microorganisms is important for the study of microbial species and for microbial applications. However, there are not many studies on ML in this area, which will be an important research direction.

Microbiome-Disease Association

There are many kinds of microorganisms in human bodies, and they are inseparable from human health. For example, intestinal microbial disorders can cause intestinal inflammatory diseases (Chen et al., 2017b), such as ulcerative colitis, CRC, atherosclerosis, diabetes and obesity. Accordingly, it is necessary to predict the microbial-disease association because this study not only improves the diagnosis and prognosis of human diseases, but also develops the new drugs (Yu et al., 2015, 2016a; Shi et al., 2016; Su et al., 2018; Fan et al., 2019). However, few studies have investigated predictive analysis of the microbial-disease association. Therefore, in this section, we introduce the application of ML to the study of microbial-disease association.

Fan et al. (2019) proposed a new approach to analyze the microbial-disease association by integrating multiple data sources from the human microbe-disease consortium (MDPH_HMDA) and path-based HeteSim scores. First, heterogeneity networks were constructed. Microbe-disease pair weighting was conducted according to the standardized HeteSim measurement method, after which the microbe-disease-disease pathway and microbe-microbe-disease pathway HeteSim scores were integrated. Finally, the correlation scores of potential micro genome associations were calculated. Xuezhong et al. (2014) proposed a method based on the Human Disease Network (HSDN) in which co-occurrence of disease/symptom terms based on PubMed bibliographic records was used to calculate disease similarity. KATZ (Katz, 1953) is a network based measurement method that calculates the similarly of nodes in a heterogeneous network, to solve the link prediction problem proposed by Katz. The KATZ method has been applied in many fields, including disease-gene association prediction (Xiaofei et al., 2014) and IncRNA-disease association prediction (Chen et al., 2015). Chen et al. (2017b) proposed a novel method based on KATZ to predict associations of human microbiota with non-infectious diseases (named KATZHMDA). The KATZHMDA first constructs adjacency matrix A based on known microbial-disease associations. The kernel similarity matrix KD and KM are calculated based on the disease Gaussian interaction profile and microbial Gaussian interaction profile, respectively. We can construct the integrated matrix A^* based on KM, KD and known microbial-disease associations. Next, all walks of different lengths are integrated to obtain a single microbe-disease association measurement. Therefore, we can calculate microbe-disease association probability in a matrix form. Shi et al. (2018) proposed a prediction method based on binary matrix completion named BMCMDA. The BMCMDA assumes that the incomplete microbiome-disease association (MDA) matrix is the sum of a potential parameterization matrix and a noise matrix. Additionally, the BMCMDA assumes that the independent subscripts of the items observed in the MDA matrix follow the binomial model. Shi et al. (2018) used the same dataset, which was collected from the Human Microbe-Disease Association Database (HMDAD) and included 292 microbes and 39 human diseases, to perform comparisons. According to the study, BMCMDA is better than the KATZHMDA in AUC. BMCMDA can be integrated with other and independent microbial/disease similarities or characteristics to enhance MDA prediction. Moreover, this method can be applied to more prediction aspects. We summarize the available datasets and methods, which are shown in **Table 3**.

TABLE 3 | The available data and materials for microbiome-disease association.

Studies	Availability of data and materials	Reference
Zhou et al.	https://www.nature.com/articles/ncomms5212#supplementary-information	Xiaofei et al., 2014
KATZHMDA	http://dwz.cn/4oX5mS.	Chen et al., 2017b
BMCMDA	https://github.com/JustinShi2016/ISBRA2017	Shi et al., 2018

CONCLUSION

Microorganisms are involved in many life activities, and affect their surrounding environment and other organisms. Microorganisms play important roles in human heath, crop growth, livestock farming, environmental management, industrial chemical production and food production. In the 19th century, people first observed microbes using microscopes and began to study them. However, the development of high-throughput sequencing technology has led to generation of large amounts of microbial related data. As a result, machine-learning methods are now being applied to microbiological research. Here, we discuss the current application of ML in the microbiome. The results revealed that ML is widely used in microbiological research, and that it has focused on classification problems and analysis of interaction problems. However, many problems remain unresolved and will require the cooperation of researchers from different fields, such as biology, informatics and medicine, to jointly promote the development and progress of microbiological research. On the other hand, the recent developed link prediction (Liu et al., 2016; Zeng et al., 2017b) and computational intelligence methods (Cabarle et al., 2017; Song et al., 2018), can be promising in discovering the relationship between diseases and microbes.

AUTHOR CONTRIBUTIONS

KQ drafted the manuscript. FG and XL conducted research. YL modified the manuscript. QZ conceived the idea.

FUNDING

The work was supported by the National Key R&D Program of China (2018YFC0910405), and the National Natural Science Foundation of China (No. 61771331).

ACKNOWLEDGMENTS

We thank Jeremy Kamen, MSc., from Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

REFERENCES

Alexander, M. (1962). Introduction of soil microbiology. *Soil Sci.* 93:74. doi: 10.1097/00010694-196201000-00034

Amgarten, D., Braga, L. P. P., da Silva, A. M., and Setubal, J. C. (2018). MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front. Genet.* 9:304. doi: 10.3389/fgene.2018.00304

Amsel, R., Totten, P. A., Spiegel, C. A., Chen, K. C., Eschenbach, D., and Holmes, K. K. (1983). Nonspecific vaginitis. Diagnostic criteria and microbial and epidemiologic associations. *Am. J. Med.* 74, 14–22. doi: 10.1016/0002-9343(83)91112-9

Asgari, E., Garakani, K., McHardy, A. C., and Mofrad, M. R. K. (2018). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics* 34, i32–i42. doi: 10.1093/bioinformatics/bty296

Atlas, R. M., and Bartha, R. (1981). Microbial ecology:fundamentals and applications. *Acta Ecol. Sin.* 70:977. doi: 10.1016/j.biortech.2015.07.074

Beck, D., and Foster, J. A. (2014). Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PLoS One* 9:e87830. doi: 10.1371/journal.pone.0087830

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., et al. (2005). Defining operational taxonomic units using DNA barcode data. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 360, 1935–1943. doi: 10.1098/rstb.2005.1725

Bourne, D. G., Garren, M., Work, T. M., Rosenberg, E., Smith, G. W., and Harvell, C. D. (2009). Microbial disease and the coral holobiont. *Trends Microbiol.* 17, 554–562. doi: 10.1016/j.tim.2009.09.004

Cabarle, F. G. C., Adorna, H. N., Jiang, M., and Zeng, X. X. (2017). Spiking neural P systems with scheduled synapses. *IEEE Trans. Nanobioscience* 16, 792–801. doi: 10.1109/tnb.2017.2762580

Chang, H. X., Haudenshield, J. S., Bowen, C. R., and Hartman, G. L. (2017). Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Front. Microbiol.* 8:519. doi: 10.3389/fmicb.2017.00519

Chen, J., Guo, M. Y., Li, S. M., and Liu, B. (2017a). ProtDec-LTR2.0: an improved method for protein remote homology detection by combining pseudo protein and supervised Learning to Rank. *Bioinformatics* 33, 3473–3476. doi: 10.1093/bioinformatics/btx429

Chen, X., Huang, Y.-A., You, Z.-H., Yan, G.-Y., and Wang, X.-S. (2017b). 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

Chen, X., Wu, Q. F., and Yan, G. Y. (2017c). RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. *RNA Biol.* 14, 952–962. doi: 10.1080/15476286.2017.1312226

Chen, X., Xie, D., Zhao, Q., and You, Z.-H. (2017d). MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 20, 515–539. doi: 10.1093/bib/bbx130

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017e). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, J., Guo, M. Y., Wang, X. L., and Liu, B. (2018a). A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinform.* 19, 231–244. doi: 10.1093/bib/bbw108

Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018b). EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.* 9:3. doi: 10.1038/s41419-017-0003-x

Chen, X., Sun, Y.-Z., Guan, N.-N., Qu, J., Huang, Z.-A., Zhu, Z.-X., et al. (2018c). Computational models for lncRNA function prediction and functional similarity calculation. *Brief. Funct. Genomics* 18, 58–82. doi: 10.1093/bfgp/ely031

Chen, X., Wang, C. C., Yin, J., and You, Z. H. (2018d). Novel human miRNA-disease association inference based on random forest. *Mol. Ther. Nucleic Acids* 13, 568–579. doi: 10.1016/j.omtn.2018.10.005

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018e). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. S. (2018f). BNPMDA: bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Yin, J., Qu, J., and Huang, L. (2018g). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, W., Ding, H., Feng, P. M., Lin, H., and Chou, K. C. (2016). IACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 7, 16895–16909. doi: 10.18632/oncotarget.7815

Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed Res. Int.* 2016:1654623. doi: 10.1155/2016/1654623

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5:11338. doi: 10.1038/srep11338

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., et al. (2017). Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. *Environ. Sci. Technol.* 51, 9118–9126. doi: 10.1021/acs.est.7b01518

Cotter, P. D., Hill, C., and Ross, R. P. (2005). Food microbiology: bacteriocins: developing innate immunity for food. *Nat. Rev. Microbiol.* 3, 777–788. doi: 10.1038/nrmicro1273

Cui, Y., Ooi, B. C., Tan, K. L., and Jagadish, H. V. (2001). "Indexing the distance: an efficient method to KNN processing", in *Vldb Proceedings of the 27th VLDB Conference*, Rome, 421–430.

Deng, Z. L., Gottschick, C., Bhuju, S., Masur, C., Abels, C., and Wagner-Dobler, I. (2018). Metatranscriptome analysis of the vaginal microbiota reveals potential mechanisms for protection against metronidazole in bacterial vaginosis. *Msphere* 3:e00262-18. doi: 10.1128/mSphereDirect.00262-18

DiMucci, D., Kon, M., and Segre, D. (2018). Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks. *Msystems* 3:e00181-18. doi: 10.1128/mSystems.00181-18

Ding, Y. J., Tang, J. J., and Guo, F. (2017a). Identification of drug-target interactions via multiple information integration. *Inf. Sci.* 418, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y. J., Tang, J. J., and Guo, F. (2017b). Identification of protein-ligand binding sites by sequence information and ensemble classifier. *J. Chem. Inf. Model.* 57, 3149–3161. doi: 10.1021/acs.jcim.7b00307

Drucker, H., Wu, D., and Vapnik, V. N. (2002). Support vector machines for spam categorization. *IEEE Trans. Neural Netw.* 10, 1048–1054. doi: 10.1109/72.788645

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

Feng, P. M., Chen, W., Lin, H., and Chou, K. C. (2013a). iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* 442, 118–125. doi: 10.1016/j.ab.2013.05.024

Feng, P. M., Ding, H., Chen, W., and Lin, H. (2013b). Naive bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* 2013:530696. doi: 10.1155/2013/530696

Feng, P. M., Lin, H., and Chen, W. (2013c). Identification of antioxidants from sequence information using naive bayes. *Comput. Math. Methods Med.* 2013:567529. doi: 10.1155/2013/567529

Feng, P. M., Ding, H., Lin, H., and Chen, W. (2017a). AOD: the antioxidant protein database. *Sci. Rep.* 7:7449. doi: 10.1038/s41598-017-08115-6

Feng, P. M., Zhang, J. D., Tang, H., Chen, W., and Lin, H. (2017b). Predicting the organelle location of noncoding RNAs using pseudo nucleotide compositions. *Interdiscip. Sci. Comput. Life Sci.* 9, 540–544. doi: 10.1007/s12539-016-0193-4

Fiannaca, A., Paglia, L. L., Rosa, M. L., Bosco, G. L., Renda, G., Rizzo, R., et al. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinformatics* 19:198. doi: 10.1186/s12859-018-2182-6

Gillevet, P., Sikaroodi, M., Keshavarzian, A., and Mutlu, E. A. (2010). Quantitative assessment of the human gut microbiome using multitag pyrosequencing. *Chem. Biodivers.* 7, 1065–1075. doi: 10.1002/cbdv.200900322

Grazziotin, A. L., Koonin, E. V., and Kristensen, D. M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* 45, D491–D498. doi: 10.1093/nar/gkw975

He, W. Y., Jia, C. Z., and Zou, Q. (2019). 4mCPred: machine learning methods for DNA N-4-methylcytosine sites prediction. *Bioinformatics* 35, 593–601. doi: 10.1093/bioinformatics/bty668

Hu, H., Zhang, L., Ai, H. X., Zhang, H., Fan, Y. T., Zhao, Q., et al. (2018). HLPI-Ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Hu, H., Zhu, C. Y., Ai, H. X., Zhang, L., Zhao, J., Zhao, Q., et al. (2017). LPI-ETSLP: lncRNA-protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction. *Mol. Biosyst.* 13, 1781–1787. doi: 10.1039/c7mb00290d

Huang, Y. A., You, Z. H., Chen, X., Huang, Z. A., Zhang, S. W., 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

Huang, Z. A., Chen, X., Zhu, Z. X., Liu, H. S., Yan, G. Y., You, Z. H., et al. (2017). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Johnson, H. R., Trinidad, D. D., Guzman, S., Khan, Z., Parziale, J. V., DeBruyn, J. M., et al. (2016). A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. *PLoS One* 11:e0167370. doi: 10.1371/journal.pone.0167370

Jolliffe, I. T. (2002). Principal component analysis. *J. Mark. Res.* 87:513.

Jordan, A. (2008). On Discriminative vs. Generative classifiers: a comparison of logistic regression and naive Bayes. *Neural Process. Lett.* 28:169. doi: 10.1007/s11063-008-9088-7

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika* 18, 39–43. doi: 10.1007/BF02289026

Kira, K., and Rendell, L. A. (1992). "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning*, Aberdeen. doi: 10.1016/B978-1-55860-247-2.50037-1

Langille, M. G. I., 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

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

Ley, R. E., Turnbaugh, P. J., Samuel, K., and Gordon, J. I. (2006b). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023.

Li, Z., Tang, J. J., and Guo, F. (2016). Learning from real imbalanced data of 14-3-3 proteins binding specificity. *Neurocomputing* 217, 83–91. doi: 10.1016/j.neucom.2016.03.093

Liao, Y., and Vemuri, V. R. (2002). Use of K-Nearest Neighbor classifier for intrusion detection. *Comput. Secur.* 21, 439–448. doi: 10.1016/S0167-4048(02)00514-X

Liu, B., Jiang, S., and Zou, Q. (2018). HITS-PR-HHblits: protein remote homology detection by combining PageRank and Hyperlink-Induced Topic Search. *Brief. Bioinform.* 2018:bby104. doi: 10.1093/bib/bby104

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K. -C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2016). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 905–915. doi: 10.1109/TCBB.2016.2550432

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145. doi: 10.1073/pnas.95.6.3140

Mainali, K. P., Bewick, S., Thielen, P., Mehoke, T., Breitwieser, F. P., Paudel, S., et al. (2017). Statistical analysis of co-occurrence patterns in microbial presence-absence datasets. *PLoS One* 12:e0187132. doi: 10.1371/journal.pone.0187132

Meena, M. J., and Chandran, K. R. (2009). "Naïve Bayes text classification with positive features selected by statistical method," in *Proceedings of the First International Conference on Advanced Computing, ICAC 2009*, Los Alamitos: IEEE, 28–33. doi: 10.1109/ICADVC.2009.5378273

Moitinho-Silva, L., Steinert, G., Nielsen, S., Hardoim, C. C. P., Wu, Y. C., McCormack, G. P., et al. (2017). Predicting the HMA-LMA status in marine sponges by machine learning. *Front. Microbiol.* 8:752. doi: 10.3389/fmicb.2017.00752

Moran, M. A. (2015). The global ocean microbiome. *Science* 350:aac8455. doi: 10.1126/science.aac8455

Morris, O. N., Cunningham, J. C., Finneycrawley, J. R., Jaques, R. P., and Kinoshita, G. (1986). Microbial insecticides in Canada: their registration and use in agriculture, forestry and public and animal health. *Bull. Entomol. Soc. Canada* 18, 1–43.

Murali, A., Bhargava, A., and Wright, E. S. (2018). IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 6:140. doi: 10.1186/s40168-018-0521-5

Nannipieri, P., Ascher, J., Ceccherini, M. T., Landi, L., Pietramellara, G., and Renella, G. (2010). Microbial diversity and soil functions. *Eur. J. Soil Sci.* 54, 655–670. doi: 10.1046/j.1351-0754.2003.0556.x

Niel, C. B. V. (1966). Microbiology and molecular biology. *Q. Rev. Biol.* 41, 105–112. doi: 10.1086/404937

Nowrousian, M. (2010). Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot. Cell* 9, 1300–1310. doi: 10.1128/EC.00123-10

Nugent, R. P., Krohn, M. A., and Hillier, S. L. (1991). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *J. Clin. Microbiol.* 29, 297–301.

Oudah, M., and Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics* 19:227. doi: 10.1186/s12859-018-2205-3

Pan, G. F., Jiang, L. M., Tang, J. J., and Guo, F. (2018). A novel computational method for detecting DNA methylation sites with DNA sequence information and physicochemical properties. *Int. J. Mol. Sci.* 19:E511. doi: 10.3390/ijms19020511

Peng, L. H., Yin, J., Zhou, L. Q., Liu, M. X., and Zhao, Y. (2018). Human microbe-disease association prediction based on adaptive boosting. *Front. Microbiol.* 9:2440. doi: 10.3389/fmicb.2018.02440

Petrof, E. O., Claud, E. C., Gloor, G. B., and Allenvercoe, E. (2012). Microbial ecosystems therapeutics: a new paradigm in medicine? *Benef. Microbes* 4, 53–65. doi: 10.3920/BM2012.0039

Podani, J., and Miklós, I. (2002). Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology* 83, 3331–3343. doi: 10.1890/0012-9658(2002)083[3331:RCATHE]2.0.CO;2

Qu, K. Y., Han, K., Wu, S., Wang, G. H., and Wei, L. Y. (2017). Identification of DNA-binding proteins using mixed feature representation methods. *Molecules* 22:E1602. doi: 10.3390/molecules22101602

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., and McCulle, S. L et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4680–4687. doi: 10.1073/pnas.1002611107

Reiff, C., and Kelly, D. (2010). Inflammatory bowel disease, gut bacteria and probiotic therapy. *Int. J. Med. Microbiol.* 300, 25–33. doi: 10.1016/j.ijmm.2009.08.004

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. Z. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69. doi: 10.1186/s40168-017-0283-5

Rodríguez, J. J., and Kuncheva, L. I. (2007). "Naïve bayes ensembles with a random oracle," in *Lecture Notes in Computer Science*, Vol. 4472, eds M. Haindl, J. Kittler and F. Roli (Berlin: Springer), 450–458.

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3:e985. doi: 10.7717/peerj.985

Schmedes, S. E., Woerner, A. E., Novroski, N. M. M., Wendt, F. R., King, J. L., Stephens, K. M., et al. (2018). Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. *Forensic Sci. Int. Genet.* 32, 50–61. doi: 10.1016/j.fsigen.2017.10.004

Schmidt, T. S. B., Matias Rodrigues, J. F., and von Mering, C. (2014). Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput. Biol.* 10:e1003594. doi: 10.1371/journal.pcbi.1003594

Shi, J. Y., Huang, H., Zhang, Y. N., Cao, J. B., and Yiu, S. M. (2018). BMCMDA: a novel model for predicting human microbe-disease associations via binary matrix completion. *BMC Bioinformatics* 19, 169–176. doi: 10.1186/s12859-018-2274-3

Shi, J. Y., Li, J. X., and Lu, H. M. (2016). Predicting existing targets for new drugs base on strategies for missing interactions. *BMC Bioinformatics* 17(Suppl. 8):282. doi: 10.1186/s12859-016-1118-2

Sibley, C. D., Parkins, M. D., Rabin, H. R., Kangmin, D., Norgaard, J. C., and Surette, M. G. (2008). A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in cystic fibrosis patients. *Proc. Natl. Acad. Sci. U.S.A.* 105, 15070–15075. doi: 10.1073/pnas.0804326105

Song, T., Rodriguez-Paion, A., Zheng, P., and Zeng, X. X. (2018). Spiking neural P systems with colored spikes. *IEEE Trans. Cogn. Dev. Syst.* 10, 1106–1115. doi: 10.1109/tcds.2017.2785332

Souza, P. M. D. (2010). Application of microbial α-amylase in industry – A review. *Braz. J. Microbiol.* 41, 850–861. doi: 10.1590/S1517-83822010000400004

Srinivasan, S., Hoffman, N. G., Morgan, M. T., Matsen, F. A., Fiedler, T. L., Hall, R. W., et al. (2012). Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS One* 7:e37818. doi: 10.1371/journal.pone.0037818

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z. G., Yang, L. Y., et al. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 1:11. doi: 10.1186/2049-2618-1-11

Stoter, F. R., Chakrabarty, S., Edler, B., and Habetse, E. A. P. (2019). CountNet: estimating the number of concurrent speakers using supervised learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* 27, 268–282. doi: 10.1109/taslp.2018.2877892

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2018). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE/ACM Trans. Comput. Biol. Bioinfor*m. doi: 10.1109/tcbb.2018.2858756 [Epub ahead of print].

Sujatha, S., Hoffman, N. G., Morgan, M. T., Matsen, F. A., Fiedler, T. L., Hall, R. W., et al. (2012). Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS One* 7:e37818. doi: 10.1371/journal.pone.0037818

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

Waldron, L. (2018). Data and statistical methods to analyze the human microbiome. *Msystems* 3:e00194-17. doi: 10.1128/mSystems.00194-17

Wang, F., Huang, Z. A., Chen, X., Zhu, Z. X., Wen, Z. K., Zhao, J. 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

Wei, L. Y., Chen, H. R., and Su, R. (2018a). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids* 12, 635–644. doi: 10.1016/j.omtn.2018.07.004

Wei, L. Y., Zhou, C., Chen, H. R., Song, J. N., and Su, R. (2018b). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Wei, L. Y., Wan, S. X., Guo, J. S., and Wong, K. K. L. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L. Y., Xing, P. W., Zeng, J. C., Chen, J. X., Su, R., and Guo, F. (2017b). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Weinbauer, M. G. (2010). Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* 28, 127–181. doi: 10.1016/j.femsre.2003.08.001

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

Wisittipanit, N. (2012). *Machine Learning Approach for Profiling Human Microbiome*. Ph.D. dissertation, George Mason University, Fairfax, VA. Available at: https://search.proquest.com/docview/1009703926?accountid=45721 (accessed April 8, 2019).

Xiaofei, Y., Lin, G., Xingli, G., Xinghua, S., Hao, W., Fei, S., et al. (2014). A network based method for analysis of lncRNA-disease associations and prediction of

lncRNAs implicated in diseases. *PLoS One* 9:e87797. doi: 10.1371/journal.pone.0087797

Xie, K., Guo, L., Bai, Y., Liu, W., Yan, J., and Bucher, M. (2018). Microbiomics and plant health: an interdisciplinary and international workshop on the plant microbiome. *Mol. Plant* 12, 1–3. doi: 10.1016/j.molp.2018.11.004

Xu, L., Liang, G. M., Liao, C. R., Chen, G. D., and Chang, C. C. (2018a). An efficient classifier for alzheimer's disease genes identification. *Molecules* 23:E3140. doi: 10.3390/molecules23123140

Xu, L., Liang, G. M., Shi, S. H., and Liao, C. R. (2018b). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Int. J. Mol. Sci.* 19:E1773. doi: 10.3390/ijms19061773

Xu, L., Liang, G. M., Wang, L. J., and Liao, C. R. (2018c). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* 9:E158. doi: 10.3390/genes9030158

Xuezhong, Z., JöRg, M., Albert-László, B., and Amitabh, S. (2014). Human symptoms-disease network. *Nat. Commun.* 5:4212. doi: 10.1038/ncomms5212

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018a). iRNA-2OM: a sequence-based predictor for identifying 2 '-o-methylation sites in homo sapiens. *J. Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004

Yang, H., Qiu, W. R., Liu, G. Q., Guo, F. B., Chen, W., Chou, K. C., et al. (2018b). iRSpot-Pse6NC: identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.* 14, 883–891. doi: 10.7150/ijbs.24616

Yang, H., Tang, H., Chen, X. X., Zhang, C. J., Zhu, P. P., Ding, H., et al. (2016). Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. *Biomed Res. Int.* 2016:5413903. doi: 10.1155/2016/5413903

Yeom, S., and Javidi, B. (2006). Automatic identification of biological microorganisms using three-dimensional complex morphology. *J. Biomed. Opt.* 11:024017.

Yu, L., Huang, J. B., Ma, Z. X., Zhang, J., Zou, Y. P., and Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomics* 8(Suppl. 2):S2. doi: 10.1186/1755-8794-8-s2-s2

Yu, L., Ma, X. K., Zhang, L., Zhang, J., and Gao, L. (2016a). Prediction of new drug indications based on clinical data and network modularity. *Sci. Rep.* 6:32530. doi: 10.1038/srep32530

Yu, L., Wang, B. B., Ma, X. K., and Gao, L. (2016b). The extraction of drug-disease correlations based on module distance in incomplete human interactome. *BMC Syst. Biol.* 10(Suppl. 4):111. doi: 10.1186/s12918-016-0364-2

Yu, L., Su, R. D., Wang, B. B., Zhang, L., Zou, Y. P., Zhang, J., et al. (2017a). Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 966–977. doi: 10.1109/tcbb.2016.2550453

Yu, L., Zhao, J., and Gao, L. (2017b). Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome. *Artif. Intell. Med.* 77, 53–63. doi: 10.1016/j.artmed.2017.03.009

Yu, L., Zhao, J., and Gao, L. (2018). Predicting potential drugs for breast cancer based on miRNA and tissue specificity. *Int. J. Biol. Sci.* 14, 971–980. doi: 10.7150/ijbs.23350

Zeng, X. X., Ding, N. X., Rodriguez-Paton, A., and Zou, Q. (2017a). Probability-based collaborative filtering model for predicting gene-disease associations. *BMC Med. Genomics* 10(Suppl. 5):76. doi: 10.1186/s12920-017-0313-y

Zeng, X. X., Lin, W., Guo, M. Z., and Zou, Q. (2017b). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1005420

Zeng, X. X., Liu, L., Lu, L. Y., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 34, 2425–2432. doi: 10.1093/bioinformatics/bty112

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. X. (2019). Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 283–291. doi: 10.1109/tcbb.2017.2776280

Zhao, Q., Liang, D., Hu, H., Ren, G. F., and Liu, H. S. (2018a). RWLPAP: random walk for lncRNA-protein associations prediction. *Protein Pept. Lett.* 25, 830–837. doi: 10.2174/0929866525666180905104904

Zhao, Q., Yu, H., Ming, Z., Hu, H., Ren, G., and Liu, H. (2018b). The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. *Mol. Ther. Nucleic Acids* 13, 464–471. doi: 10.1016/j.omtn.2018.09.020

Zhao, Q., Zhang, Y., Hu, H., Ren, G. F., Zhang, W., and Liu, H. S. (2018c). IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Int. J. Inf. Fusion* 50, 71–91. doi: 10.1016/j.inffus.2018.09.012

Zou, Q., Chen, L., Huang, T., Zhang, Z.G., and Xu, Y.G. (2017). Machine learning and graph analytics in computational biomedicine. *Artif. Intell. Med.* 83:1. doi: 10.1016/j.artmed.2017.09.003

Zou, Q., Li, J. J., Song, L., Zeng, X. X., and Wang, G. H. (2016). Similarity computation strategies in the microRNA-disease network: a survey. *Brief. Funct. Genomics* 15, 55–64. doi: 10.1093/bfgp/elv024

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018a). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* bby090. doi: 10.1093/bib/bby090

Zou, Q., Qu, K. Y., Luo, Y. M., Yin, D. H., Ju, Y., and Tang, H. (2018b). Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* 9:515. doi: 10.3389/fgene.2018.00515

Conflict of Interest Statement: 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.

Copyright © 2019 Qu, Guo, Liu, Lin and Zou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 April 2019
doi: 10.3389/fmicb.2019.00826

Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model

Dongmei Ai[1,2]*, *Hongfei Pan*[2], *Xiaoxin Li*[2], *Yingxin Gao*[2], *Gang Liu*[2] and *Li C. Xia*[3]*

[1] Basic Experimental of Natural Science, University of Science and Technology Beijing, Beijing, China, [2] School of Mathematics and Physics, University of Science and Technology Beijing, Beijing, China, [3] Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

Colorectal cancer (CRC) is the third most common cancer worldwide. Its incidence is still increasing, and the mortality rate is high. New therapeutic and prognostic strategies are urgently needed. It became increasingly recognized that the gut microbiota composition differs significantly between healthy people and CRC patients. Thus, identifying the difference between gut microbiota of the healthy people and CRC patients is fundamental to understand these microbes' functional roles in the development of CRC. We studied the microbial community structure of a CRC metagenomic dataset of 156 patients and healthy controls, and analyzed the diversity, differentially abundant bacteria, and co-occurrence networks. We applied a modified zero-inflated lognormal (ZIL) model for estimating the relative abundance. We found that the abundance of genera: *Anaerostipes, Bilophila, Catenibacterium, Coprococcus, Desulfovibrio, Flavonifractor, Porphyromonas*, *Pseudoflavonifractor,* and *Weissella* was significantly different between the healthy and CRC groups. We also found that bacteria such as *Streptococcus, Parvimonas, Collinsella, and Citrobacter* were uniquely co-occurring within the CRC patients. In addition, we found that the microbial diversity of healthy controls is significantly higher than that of the CRC patients, which indicated a significant negative correlation between gut microbiota diversity and the stage of CRC. Collectively, our results strengthened the view that individual microbes as well as the overall structure of gut microbiota were co-evolving with CRC.

Keywords: gut microbiota, colorectal cancer, zero-inflated lognormal model, association network, microbial diversity

OPEN ACCESS

Edited by:
Qi Zhao,
Liaoning University, China

Reviewed by:
Quan Chen,
Icahn School of Medicine at Mount Sinai, United States
Lingling An,
University of Arizona, United States

***Correspondence:**
Dongmei Ai
aidongmei@ustb.edu.cn
Li C. Xia
l.c.xia@stanford.edu

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

Received: *15 January 2019*
Accepted: *01 April 2019*
Published: *24 April 2019*

Citation:
Ai D, Pan H, Li X, Gao Y, Liu G and Xia LC (2019) Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model. Front. Microbiol. 10:826. doi: 10.3389/fmicb.2019.00826

INTRODUCTION

A large number of microbes colonize the human body. They form a complex microbial community, or microbiota (Tringe et al., 2005; Zhao et al., 2013; Liao et al., 2015). Among them, the gut microbiota is the most diverse, with more than 1,000 species (Kostic et al., 2012; Li et al., 2012; Ahn et al., 2013). Those microbes are involved in maintaining intestinal homeostasis, through physiological processes such as metabolism, immune responses, and inflammation, all of which are essential for human health. Previous studies revealed a deliciated and dynamic balance between the microbial community and the host, which is likely the result of long term co-evolution. However,

studies also observed that pathogenic changes in the structure, composition, and function of gut microbiota can lead to various diseases, often by causing the production of abnormal metabolites (Chen et al., 2016a; Huang et al., 2017a,b). Those diseases and conditions include irritable bowel syndrome (Kipanyula et al., 2013), Crohn's disease (Sommer and Bäckhed, 2013), and colorectal cancer (CRC) (Zackular et al., 2014; Rea et al., 2018).

The mechanisms by which gut microbes influence the CRC tumorigenesis (Iacob et al., 2017) were actively under study. For examples, researchers have recently learned that the gut microbiota plays a regulatory role in the tumor microenvironment and thus in tissue carcinogenesis (Sohn et al., 2015; Nagy-Szakal et al., 2017; Morgillo et al., 2018). Guo et al. also found that the microbiota structure and microbial metabolites can affect the body's susceptibility to CRC by directly inducing pathological conditions, such as adenoma (Guo et al., 2015). However, to further understand such interactions, it is essential to characterize and compare the gut microbiota structure of healthy controls and cancer patients. And based on that, specific microbiota patterns or strain types need to be identified to provide new targets and strategies for cancer prevention and treatment (Hu et al., 2017, 2018; Zhao et al., 2018a,b,c). Therefore, in this paper, we aim to determine the microbes that are associated with CRC using a large-scale metagenomic data set.

While the metagenomics research has provided enormous scientific data for investigating the role of the gut microbiota in the context of cancer development and progression (Zhang et al., 2014), appropriate bioinformatics and statistical analyses are also required to accurately identifying the differential microbes. Several algorithms using either parametric or non-parametric tests have been proposed to determine such species. For examples, Abusleme et al. (2013) combined the Kruskal-Wallis test with the Wilcoxon rank-sum test to analyze periodontitis data and used linear discriminant analysis to identify the species with significant differences between periodontitis patients and healthy controls. Nagy-Szakal et al. used the non-parametric Mann-Whitney *U* test with Benjamini-Hochberg correction to show that the microbial composition in the intestines of patients with chronic fatigue syndrome differed significantly from that of healthy individuals (Nagy-Szakal et al., 2017). And Peng et al. conducted beta regression on the abundance of microbes to obtain regression coefficients (Peng et al., 2016).

One particular difficulty associated with the statistical testing of differential abundance is the under-sampling or dropout (Hughes et al., 2001) of less abundant microbes caused by an insufficient sequencing depth. This fact creates many zeros in the abundance values and leads to inaccurate differential analysis when only conventional normalization was applied. This issue might be mitigated with the Zero-inflated Negative Binomial modeling (ZINB) (Ridout et al., 1998). The method is now widely adopted. For examples, Paulson et al. analyzed the differential abundance in sparse high-throughput large-scale microbial marker gene survey data by using a zero-inflated Gaussian distribution mixture model with cumulative-sum scaling normalization (Paulson et al., 2013). Zhang et al. (2016) identified differentially abundant taxa between two or more populations by using a ZINB regression method and estimated the model parameters by Expectation Maximization algorithm. Chen et al. proposed a zero-inflated Beta regression model which included two parts: a logistic regression component and a Beta regression component, for testing the association between microbial abundance and clinical covariates for longitudinal microbiome data (Chen and Li, 2016). Chen Jun et al. in 2017, proposed a robust and powerful framework of differential analysis of microbiome data based on a zero-inflated negative binomial (ZINB) regression model (Chen et al., 2017). They also proposed an omnibus test of all the parameters. Omnibus test was compared with previous methods [edgeR (Robinson et al., 2010), RAIDA (Sohn et al., 2015), DESeq2 (Love et al., 2014), and metagenomeSeq (Paulson et al., 2013)] by using simulated data. RAIDA had slightly worse FDR control at a high nominal level than omnibus test, but better FDR control than other methods. The performance of RAIDA was close to that of the omnibus test, and were higher than one of other methods. RAIDA is more effective at controlling FPR than other method including the omnibus test.

In this study, we identified the differentially abundant gut microbes between CRC and healthy samples using the *Ratio Approach for Identifying Differential Abundance* (RAIDA) algorithm (Sohn et al., 2015). The algorithm fitted the distribution of observed data with a modified zero-inflated lognormal (ZIL) model and estimated the statistical significance of abundance difference by the *T*-test. Furthermore, we used the GRAMMy algorithm (Xia et al., 2011) to estimate and analyze the relative abundance of gut microbes and diversity of the microbial communities. Finally, we constructed and analyzed a microbial association network based on all healthy, small adenoma, large adenoma, and CRC samples.

MATERIALS AND METHODS

Two Metagenomics Datasets

Our first gut metagenomics dataset was downloaded from the European Nucleotide Archive (ENA) database (accession number ERP005534) (**Table 1**). The dataset (Zeller et al., 2014) consists of 156 samples from France (61 healthy, 27 small adenoma, 15 large adenoma, and 53 CRC samples). Samples with an adenoma diameter smaller than 10 mm were classified as small adenoma while those with larger than 10 mm ones were classified as large adenoma.

Our second gut metagenomics dataset was also downloaded from the ENA database (accession number ERP008729) (Zeller et al., 2014). The dataset included 156 samples from Austria, including 63 healthy samples, 47 adenoma patient samples, and 46 CRC patient samples.

A Modified ZIL Model

We estimated the relative abundance of gut microbes using the GRAMMy algorithm. We then identified differentially abundant microbes by the RAIDA algorithm which uses a modified ZIL model to account for ratios with zeros. Metagenomic data are typically sparse because of undersampling

TABLE 1 | Number of experimental samples.

Total number of samples	Healthy control	Adenoma		Colorectal cancer			
		Small (<1 cm)	Large (>1 cm)	Early stage		Late stage	
				I	II	III	IV
156	61	27	15	15	7	10	21

of the microbial community or insufficient sequencing depth. The resulting abundance table is over-presented with zeros assumed that most of those zeros is a result of insufficient sequencing depth, i.e., the under-sampling of the microbial community. Based on the assumption that most microbes are not differentially abundant, the RAIDA algorithm was systematically demonstrated to consistently identify differentially abundant microbes. We adapted the RAIDA model for our statistical analysis as follows.

Let γ_{ij} denote the observed count for microbes i and sample j, and let r_{ij} denote the ratio of γ_{ij} to γ_{kj}, where k represents the microbe (or a set of microbes) used as a divisor and $\gamma_{kj} > 0$ for all j. Here, $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$. The abundance ratio computed this way is denoted as R_{ij}^{ε} such that:

$$R_{ij}^{\varepsilon} \sim \begin{cases} Unif(0, \varepsilon) & with\ probability\ p_i \\ LN(\mu_i, \sigma_i^2) & with\ probability\ 1 - p_i \end{cases} \quad (1)$$

In this study, we used $\varepsilon = \min(r_{ij} \,|\, r_{ij} > 0)$ for all i and j. The parameters $\theta_i = (\alpha_i, \mu_i, \sigma_i)$ were estimated by the following expectation-maximization (EM) algorithm. Given that a ratio R follows a lognormal distribution, thus:

$$LN(r \,|\, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}\, r} \exp\left[-\frac{(\log r - \mu)^2}{2\sigma^2}\right], \quad (2)$$

in which, by definition, $Y = \log R$ is normally distributed with mean μ and variance σ^2. Let $y_{ij} = \log r_{ij}^{\varepsilon}$, z_{ij} is an unobservable latent variable that accounts for the probability of zero coming from the false state. Thus, the maximum-likelihood estimate of θ_i for the modified ZIL model, i.e., Equation (1), can be obtained by solving

$$\begin{aligned} \ell(\theta_i \,|\, y_{ij}, z_{ij}) = & \sum_{j=1}^{m} z_{ij} \log\left[\eta_i + (1 - p_i)\phi(y_{ij};\, \mu_i, \sigma_i^2)\right] \\ & + \sum_{j=1}^{m} (1 - z_{ij}) \log(1 - p_i) \\ & + \sum_{j=1}^{m} (1 - z_{ij}) \log \phi(y_{ij};\, \mu_i, \sigma_i^2), \end{aligned} \quad (3)$$

where ϕ is the probability density function of a normal distribution.

Diversity Analysis

To analyze microbial diversity, alpha diversity was used to measure the differences in gut microbial structure in the following three stages: healthy, adenoma (small and large combined), and cancer. We used the Shannon diversity index to measure the alpha diversity of the gut community. The Shannon index is defined as

$$H = -\sum_{j=1}^{N} a_j \ln a_j, \quad (4)$$

where H represents the Shannon Index, N indicates the total number of microbial species detected, and a_j indicates the relative abundance of the j th microorganism.

RESULTS AND DISCUSSION

Alpha Diversity of Gut Microbiota Predicts Colorectal Cancer Status

We computed the alpha diversity of gut microbes of the healthy samples, adenoma samples and CRC samples using the Shannon index and compared them with the rank-sum Dunn test (**Figure 1**). We found that the alpha diversity was significantly lower in the CRC samples as compared to the healthy samples (two tailed, Dunn test, $P < 0.0001$) and adenoma samples (two tailed, Dunn test, $P = 0.0021$). However, the alpha diversity of the healthy and adenoma samples was not significantly different (two tailed, Dunn test, $P = 0.0571$). To study the relationship between the probability of cancer occurrence and the alpha diversity, we performed logit regression to associate CRC status with the Shannon index. The regression results showed that the Shannon index is a significant predictor of CRC status (univariate

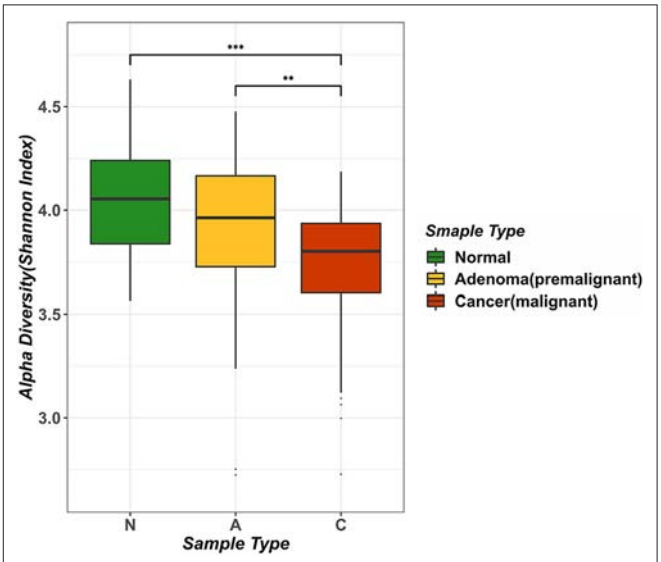


FIGURE 1 | Analysis of intestinal microbial diversity in different environments. The three colors in the figure indicate the microbial diversity in different states: green represents the healthy samples, yellow represents adenoma (precancerous lesion) growth in the intestine, and red represents a sample of colorectal cancer patients. The average value of Alpha diversity of healthy samples was 4.0456, whereas the counterpart in the adenoma sample was 3.8957, and that in the cancer sample was 3.7161.

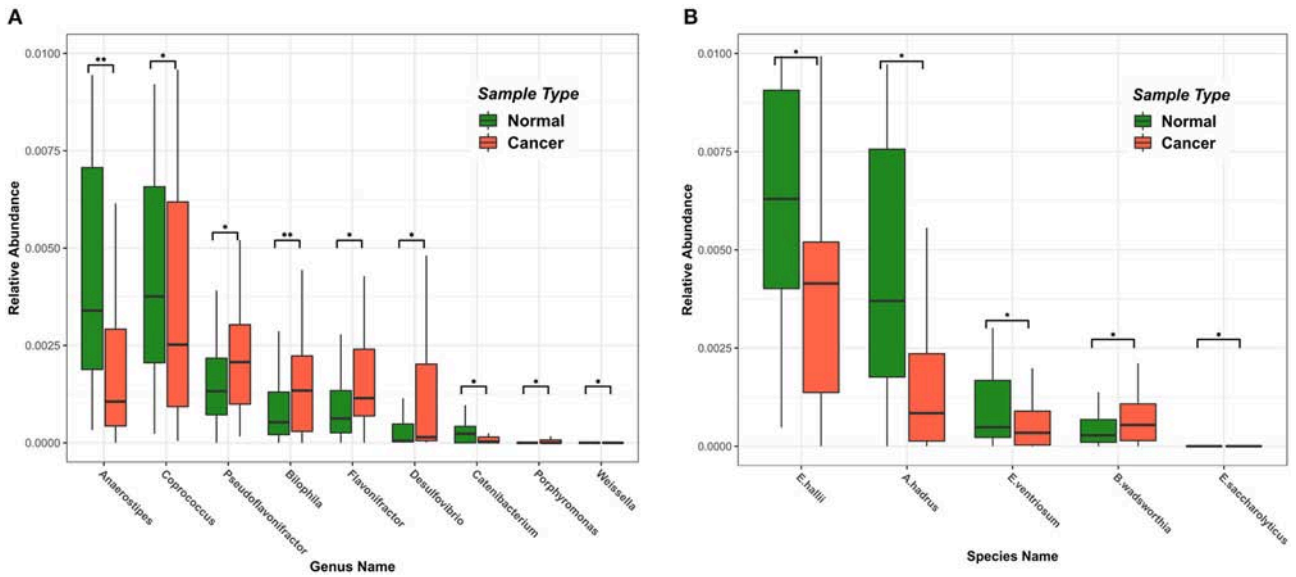


FIGURE 2 | Differences in intestinal microbiome at the genus and species level among samples of different states. Green represents a healthy sample, and red represents a colorectal cancer sample. **(A)** The top nine microbial genera with significant differences in abundance. **(B)** The top five microbial strains with significant differences in abundance.

logistic model, $P < 0.05$). The fitted logistic regression model was as follows:

$$P = \frac{\exp(-4.563d + 17.546)}{1 + \exp(-4.563d + 17.546)}, \quad (5)$$

i.e., $\text{logit}(P) = -4.563d + 17.546$, where P is the probability of being CRC, and d is the Shannon diversity index. We provided the plot of the relationship of probability of cancer occurrence and Shannon index of adenoma patients as show in **Figure S1**. Our result suggested that the diversity of the microbial species in the human intestines decreases as colorectal malignancies grow, which was supported by literature (Ahn et al., 2013).

Nine Genera Were Differentially Abundant in the Colorectal Cancer Gut Environment

Using the RAIDA algorithm, we identified nine microbial genera that were significantly different in abundance between the CRC and the controls, which included *Anaerostipes, Coprococcus, Pseudoflavonifractor, Bilophila, Flavonifractor, Desulfovibrio, Catenibacterium, Porphyromonas*, and *Weissella* (**Figure 2A**). We first observed that the abundance of *Coprococcus* was higher in the healthy samples as compared to the CRC patients. As a validation, Shen et al. showed that colorectal adenomas had lower relative abundance of *Bacteroides* spp. and *Coprococcus* spp. than controls (Shen et al., 2010). The metabolic activity of butyrate-producing bacteria is the major source of butyrate in human body. *Coprococcus* is among the essential butyrate-producing genera in human body, which promote colonic health by mediating anti-inflammatory and antitumor effects, as well as providing energy for colonocytes (Singh et al., 2014).

Also notable in our result were the genera *Fusobacterium (Fusobacteriaceae)* and *Porphyromonas (Porphyromonadaceae)*, which were shown highly enriched in the CRC patients. So was the species *Bibliophile wadsworthia*. Those sulfidogenic bacteria, including *Desulfovibrio*, *Fusobacterium,* and *Bilophila wadsworthia*, likely participate in the development of CRC by producing hydrogen sulfide (Ridlon et al., 2016; Dahmus et al., 2018). *Bilophila wadsworthia* was additionally reported to cause systemic inflammatory response in a preclinical mice study (Zhou et al., 2017).

Interestingly, we also observed that the abundance of *Eubacterium hallii, Anaerostipes hadrus,* and *Eubacterium ventriosum* (**Figure 2B**) were significantly higher in the healthy samples than in the CRC samples. *E. hallii* and *A. hadrus* can utilize the glucose and fermentation intermediates acetate and lactate to form butyrate and hydrogen, which were considered important microbes in maintaining intestinal metabolic balance (Christina et al., 2016).

We also found that *Flavonifractor* was higher in the healthy samples than that in the CRC samples, which was in agreement with Anand et al. (2016). We also observed that *Anaerostipes* had a significantly lower abundance in the CRC samples, which agreed with previous studies (Peters et al., 2016; Mori et al., 2018). We found that no *Catenibacterium* and *Gardnerella* (*Bifidobacteriaceae*) were present in CRC patient samples, which was supported by Chen et al. (2012).

We tested if the nine differentially abundant genera are viable biomarkers to distinguish healthy individuals from CRC patients.

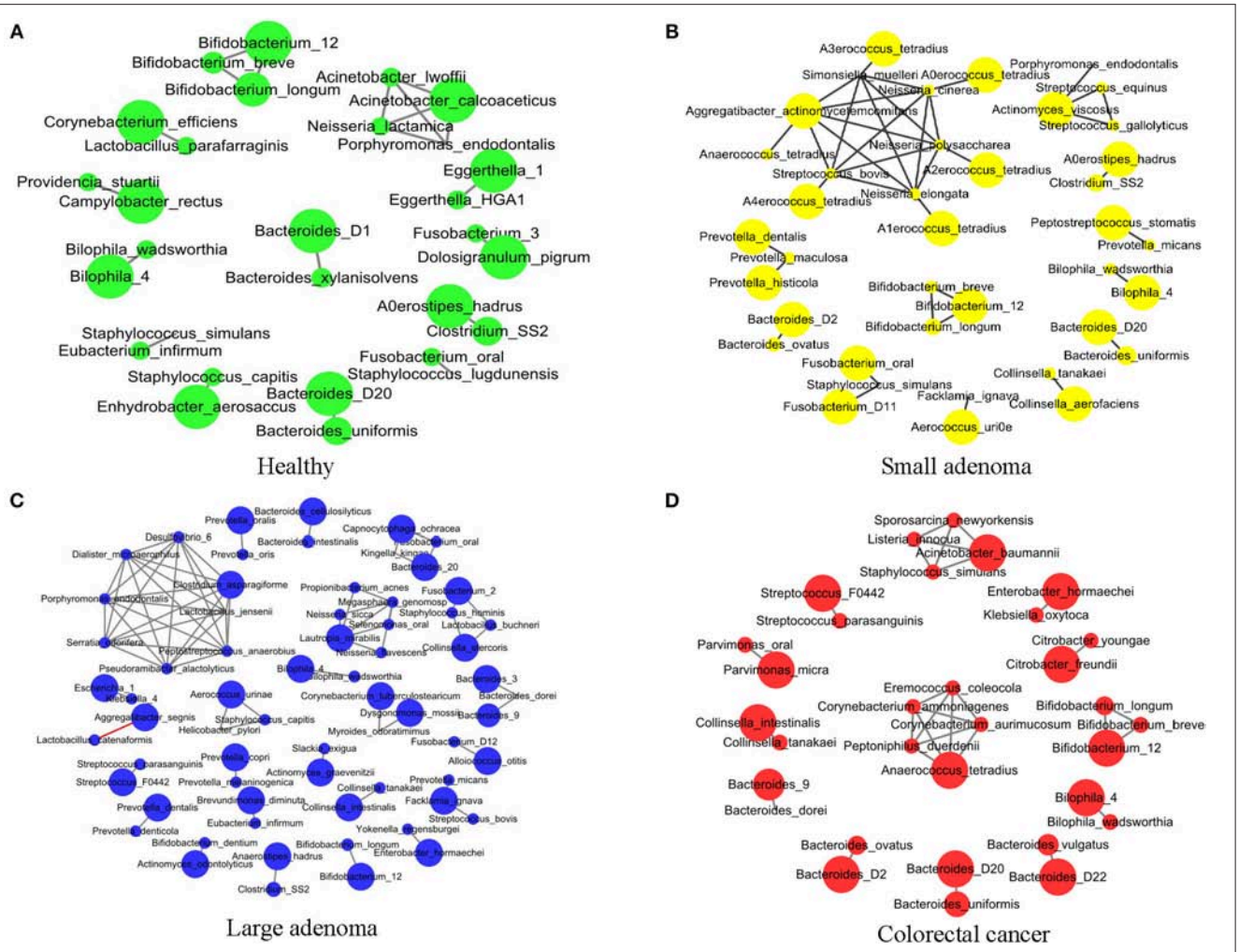


FIGURE 3 | The association network of intestinal microbiome in different states. Each circle represents the average relative abundance of a microbial species in that state. The higher the average relative abundance, the larger the area of the circle. The solid gray line between the circles indicates a positive Spearman correlation between the two groups, and the solid red line indicates a negative Spearman correlation between the two groups. **(A–D)** The association network of intestinal microbiome in healthy, small adenoma, large adenoma and cancer samples.

We trained a random forest classifier using a 5-fold cross-validation (rotative using 80% data as the training set the rest 20% as the testing set) using the first metagenomic dataset. The classifier achieved an Area Under Curve (AUC) of 0.9333.

Microbial Co-occurrence Network Evolves With CRC Development

Sophie Weiss et al. compared 8 methods of establishing association networks, they recommend filtering out extremely rare OTUs prior to network construction (Weiss et al., 2016). According to Figure 7 in this paper, SparCC should be used when the inverse simpson n_{eff} of microbes < 13, SparCC maintain high precision compared with predictions on abundance tables with low n_{eff}. But the inverse simpson n_{eff} of microbes is 27.9 (>13) in our paper, abundance of OTUs are more than 50% sparse. So we calculated the correlation between species by Pearson correlation coefficient (Pearson, 1909). We further conducted an association network analysis to identify the co-occurring intestinal microbes under different CRC states. All significant co-occurrences (PCC > 0.5) were found to be within the same genera, such as *Bifidobacterium*, *Bacteroides*, and *Bilophila* (**Figure 3**). Furthermore, both *Bifidobacterium* and *Bacteroides* were previously identified by us to have significant differences in abundance between healthy controls and CRC patients (**Figure 3A**). It is thus reasonable to assess that these bacteria were pathogenic as a group because the change of abundance in one them can result in changes of abundance in the entire clique. Our observation supported the theory that CRC ensues an interrupted balance between these bacteria (Brennan and Garrett, 2016; Yazici et al., 2017).

Co-occurrence was also found among species of the genus *Prevotella* in the healthy, small adenoma, and large adenoma environments (**Figures 3A–C**), however, such co-occurrence was missing in the CRC environment (**Figure 3D**). Conversely, several species of the genera *Streptococcus, Parvimonas, Collinsella,* and *Citrobacter* were only co-occurring in the cancer environment. Overall, we observed fewer microbial co-occurrences the healthy environment. While, in the adenoma environments, we found an increase of co-occurring pathogenic microbes. The number of co-occurring microbes was then reduced in the CRC environment. The total number of co-occurrence is relatively close between the healthy and the CRC environment, however, the microbes involved were distinct. The number of total co-occurrence might have peaked at the adenoma environments because of the co-existence of competing homeostatic and pathogenic microbial interactions in the intermediacy stage.

CONCLUSIONS

We analyzed the alpha diversity of the gut microbial community of 156 healthy, adenoma and CRC samples. We found the alpha diversity was significantly higher in healthy samples as compared to the CRC samples. We applied a modified ZIL model and identified nine significantly different genera between the healthy and CRC groups, i.e., *Anaerostipes, Bilophila, Catenibacterium, Coprococcus, Desulfovibrio, Flavonifractor, Porphyromonas, Pseudoflavonifractor,* and *Weissella*. We used these nine genera as input features for a random forest classifier and successfully predicted the CRC status with a high AUC score of 0.9333. Our results suggested that the community member and the overall structure of the gut microbiota are potential effective biomarkers of CRC stages. This avenue is being actively pursued by us and other computational researchers (Chen and Yan, 2013; Chen et al., 2016b,c, 2018a,b,c; Chen and Huang, 2017), who may bring in novel strategies for preventing and curing CRC in the near future.

AUTHOR CONTRIBUTIONS

DA and YG conducted the analysis, summarized the result and drafted the manuscript. HP, XL, and GL assisted in the data analysis and contributed to the manuscript. DA and LX conceived the study. LX supervised the manuscript writing. All authors have read and approved the final manuscript.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (61873027, 61370131). LX is supported by the Innovation in Cancer Informatics Fund, American Cancer Society (132922-PF-18-184-31301-TBG), the National Institutes of Health (HG006137-07), and funds from the Intermountain Healthcare.

ACKNOWLEDGMENTS

Both DA and LX thank Professor Fengzhu Sun at the University of Southern California and LX thanks Dr. Nancy Zhang at the University of Pennsylvania and Dr. Hanlee Ji at Stanford University for their support and helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00826/full#supplementary-material

Figure S1 | Logit regression prediction results of the Shannon diversity index. The blue circle in the figure represents a large adenoma sample, and the red triangle represents a small adenoma sample.

REFERENCES

Abusleme, L., Dupuy, A. K., Dutzan, N., Silva, N., Burleson, J. A., Strausbaugh, L. D., et al. (2013). The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *ISME J.* 7, 1016–1025. doi: 10.1038/ismej.2012.174

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

Anand, S., Kaur, H., and Mande, S. S. (2016). Comparative *in silico* analysis of butyrate production pathways in gut commensals and pathogens. *Front. Microbiol.* 7:1945. doi: 10.3389/fmicb.2016.01945

Brennan, C. A., and Garrett, W. S. (2016). Gut microbiota, inflammation, and colorectal cancer. *Annu. Rev. Microbiol.* 70, 395–411. doi: 10.1146/annurev-micro-102215-095513

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

Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., et al. (2017). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics* 34, 643–651. doi: 10.1093/bioinformatics/btx650

Chen, W., Liu, F., Ling, Z., Tong, X., and Xiang, C. (2012). Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS ONE* 7:e39743. doi: 10.1371/journal.pone.0039743

Chen, X., and Huang, L. (2017). LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2016a). 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

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016b). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput. Biol.* 12:e1004975. doi: 10.1371/journal.pcbi.1004975

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018a). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.-H., and Liu, H. (2018b). BNPMDA: bipartite network projection for MiRNA–disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2016c). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinformatics* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., and Yan, G.-Y. (2013). Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Christina, E., Hans-Joachim, R., Niko, B., Christophe, L., and Clarissa, S. (2016). The common gut microbe Eubacterium halliialso contributes to intestinal propionate formation. *Front. Microbiol.* 7:713. doi: 10.3389/fmicb.2016.00713

Dahmus, J. D., Kotler, D. L., Kastenberg, D. M., and Kistler, C. A. (2018). The gut microbiome and colorectal cancer: a review of bacterial pathogenesis. *J. Gastrointest. Oncol.* 9, 769–777. doi: 10.21037/jgo.2018.04.07

Guo, H., Shao, Y., Menghe, B., and Zhang, H. (2015). Research on the relation between gastrointestinal microbiota and disease. *Microbiol. China* 42, 400–410. doi: 10.13344/j.microbiol.china.140474

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Hu, H., Zhu, C., Ai, H., Zhang, L., Zhao, J., Zhao, Q., et al. (2017). LPI-ETSLP: lncRNA–protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction. *Mol. Biosyst.* 13, 1781–1787. doi: 10.1039/C7MB00290D

Huang, Y. A., You, Z. H., Chen, X., Huang, Z. A., Zhang, S., and Yan, G. Y. (2017a). 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

Huang, Z. A., Chen, X., Zhu, Z., Liu, H., Yan, G. Y., You, Z. H., et al. (2017b). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., and Bohannan, B. J. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* 67, 4399–4406. doi: 10.1128/AEM.67.10.4399-4406.2001

Iacob, T., Tătulescu, D. F., and Dumitraşcu, D. (2017). Therapy of the postinfectious irritable bowel syndrome: an update. *Clujul Med.* 90, 133–138. doi: 10.15386/cjmed-752

Kipanyula, M. J., Etet, P. F. S., Vecchio, L., Farahna, M., Nukenine, E. N., and Kamdje, A. H. N. (2013). Signaling pathways bridging microbial-triggered inflammation and cancer. *Cell. Signal.* 25, 403–416. doi: 10.1016/j.cellsig.2012.10.014

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

Li, Q., Wang, C., Tang, C., Li, N., and Li, J. (2012). Molecular-phylogenetic characterization of the microbiota in ulcerated and non-ulcerated regions in the patients with Crohn's disease. *PLoS ONE* 7:e34939. doi: 10.1371/journal.pone.0034939

Liao, B., Wang, S., Zhang, J., and Hongjing, Y. U. (2015). Role of gut microbiota in human diseases. *Chin. J. Gastroenterol.* 20, 126–128. doi: 10.3969/j.issn.1008-7125.2015.02.015

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Morgillo, F., Dallio, M., Della Corte, C. M., Gravina, A. G., Viscardi, G., Loguercio, C., et al. (2018). Carcinogenesis as a result of multiple inflammatory and oxidative hits: a comprehensive review from tumor microenvironment to gut microbiota. *Neoplasia* 20, 721–733. doi: 10.1016/j.neo.2018.05.002

Mori, G., Rampelli, S., Orena, B. S., Rengucci, C., De Maio, G., Barbieri, G., et al. (2018). Shifts of faecal microbiota during sporadic colorectal carcinogenesis. *Sci. Rep.* 8:10329. doi: 10.1038/s41598-018-28671-9

Nagy-Szakal, D., Williams, B. L., Mishra, N., Che, X., Lee, B., Bateman, L., et al. (2017). Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome* 5:44. doi: 10.1186/s40168-017-0261-y

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

Pearson, K. (1909). Determination of the coefficient of correlation. *Science* 30, 23–25. doi: 10.1126/science.30.757.23

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

Peters, B. A., Dominianni, C., Shapiro, J. A., Church, T. R., Wu, J., Miller, G., et al. (2016). The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome* 4:69. doi: 10.1186/s40168-016-0218-6

Rea, D., Coppola, G., Palma, G., Barbieri, A., Luciano, A., Del Prete, P., et al. (2018). Microbiota effects on cancer: from risks to therapies. *Oncotarget* 9:17915. doi: 10.18632/oncotarget.24681

Ridlon, J. M., Wolf, P. G., and Gaskins, H. R. (2016). Taurocholic acid metabolism by gut microbes and colon cancer. *Gut Microbes* 7, 201–215. doi: 10.1080/19490976.2016.1150414

Ridout, M., Demétrio, C. G., and Hinde, J. (1998). "Models for count data with many zeros," in *Proceedings of the XIXth International Biometric Conference: International Biometric Society Invited Papers* (Cape Town), 179–192.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Shen, X. J., Rawls, J. F., Randall, T., Burcal, L., Mpande, C. N., Jenkins, N., et al. (2010). Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. *Gut Microbes* 1, 138–147. doi: 10.4161/gmic.1.3.12360

Singh, N., Gurav, A., Sivaprakasam, S., Brady, E., Padia, R., Shi, H., et al. (2014). Activation of Gpr109a, receptor for niacin and the commensal metabolite butyrate, suppresses colonic inflammation and carcinogenesis. *Immunity* 40, 128–139. doi: 10.1016/j.immuni.2013.12.007

Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics* 31, 2269–2275. doi: 10.1093/bioinformatics/btv165

Sommer, F., and Bäckhed, F. (2013). The gut microbiota—masters of host development and physiology. *Nat. Rev. Microbiol.* 11, 227–238. doi: 10.1038/nrmicro2974

Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., et al. (2005). Comparative metagenomics of microbial communities. *Science* 308, 554–557. doi: 10.1126/science.1107851

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, 1669–1681. doi: 10.1038/ismej.2015.235

Xia, L. C., Cram, J. A., Chen, T., Fuhrman, J. A., and Sun, F. (2011). Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS ONE* 6:e27992. doi: 10.1371/journal.pone.0027992

Yazici, C., Wolf, P. G., Kim, H., Cross, T.-W. L., Vermillion, K., Carroll, T., et al. (2017). Race-dependent association of sulfidogenic bacteria with colorectal cancer. *Gut* 66, 1983–1994. doi: 10.1136/gutjnl-2016-313321

Zackular, J. P., Rogers, M. A., Ruffin, M. T., and Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res.* 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

Zhang, X., Mallick, H., and Yi, N. (2016). Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. *J. Bioinform. Genom.* 2, 1–9. doi: 10.18454/jbg.2016.2.2.1

Zhang, Z., Liu, C. H., and Zhao, X. H. (2014). Research advance of human gut microbiome and related diseases. *Chin. Bull. Life Sci.* 26, 768–772. doi: 10.13376/j.cbls/2014108

Zhao, Q., Liang, D., Hu, H., Ren, G., and Liu, H. (2018a). RWLPAP: random walk for lncRNA-protein associations prediction. *Protein Pept. Lett.* 25, 830–837. doi: 10.2174/0929866525666180905104904

Zhao, Q., Yu, H., Ming, Z., Hu, H., Ren, G., and Liu, H. (2018b). The Bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. *Mol. Ther.* 13, 464–471. doi: 10.1016/j.omtn.2018.09.020

Zhao, Q., Yue, Z., Hu, H., Ren, G., Wen, Z., and Liu, H. (2018c). IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zhao, Y., Wu, J., Li, J. V., Zhou, N.-Y., Tang, H., and Wang, Y. (2013). Gut microbiota composition modifies fecal metabolic profiles in mice. *J. Proteome Res.* 12, 2987–2999. doi: 10.1021/pr400263n

Zhou, F., Long, W., Hao, B., Ding, D., Ma, X., Zhao, L., et al. (2017). A human stool-derived Bilophila wadsworthia strain caused systemic inflammation in specific-pathogen-free mice. *Gut Pathog.* 9:59. doi: 10.1186/s13099-017-0208-7

Conflict of Interest Statement: 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.

Copyright © 2019 Ai, Pan, Li, Gao, Liu and Xia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

METHODS
published: 16 May 2019
doi: 10.3389/fmicb.2019.01084

Evaluating the Effect of QIIME Balanced Default Parameters on Metataxonomic Analysis Workflows With a Mock Community

*Dimitrios Kioroglou, Albert Mas and Maria del Carmen Portillo**

Department Biochemistry and Biotechnology, Faculty of Oenology, University Rovira i Virgili, Tarragona, Spain

Metataxonomic analysis represents a fast and cost-effective approach for acquiring informative insight into the composition of the microbiome of samples with variable diversity, such as wine samples. Nevertheless, it comprises a vast amount of laboratory procedures and bioinformatic frameworks each one associated with an inherent variability of protocols and algorithms, respectively. As a solution to the bioinformatic maze, QIIME bioinformatic framework has incorporated benchmarked, and balanced parameters as default parameters. In the current study, metataxonomic analysis of two types of mock community standards with the same microbial composition has been performed for evaluating the effectivess of QIIME balanced default parameters on a variety of aspects related to different laboratory and bioinformatic workflows. These aspects concern NGS platforms, PCR protocols, bioinformatic pipelines, and taxonomic classification algorithms. Several qualitative performance expectations have been the outcome of the analysis, rendering the mock community a useful evaluation tool.

Keywords: metataxonomics, next-generation-sequencing, bioinformatics, QIIME, PCR, Ion Torrent, Illumina, wine

OPEN ACCESS

Edited by:
Qi Zhao,
Shenyang Aerospace University, China

Reviewed by:
Keith A. Crandall,
George Washington University, United States
Graziano Pesole,
University of Bari Aldo Moro, Italy

****Correspondence:***
Maria del Carmen Portillo
carmen.portillo@urv.cat

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

Received: *25 February 2019*
Accepted: *29 April 2019*
Published: *16 May 2019*

Citation:
Kioroglou D, Mas A and Portillo MC (2019) Evaluating the Effect of QIIME Balanced Default Parameters on Metataxonomic Analysis Workflows With a Mock Community. Front. Microbiol. 10:1084. doi: 10.3389/fmicb.2019.01084

1. INTRODUCTION

During the past years significant improvements in Next Generation Sequencing (NGS) platforms and computational performance have given a considerable momentum to the research of microbial communities. Primarily there are two sequencing-based methods for the classification analysis of a microbiome, the metagenomic approach which concerns the shotgun sequencing of microbial DNA, and the metataxonomic approach which refers to the sequencing of a marker gene, having as a usual target the ribosomal RNA gene (Breitwieser et al., 2017). Due to the cost-effectiveness and decreased demands on computational resources of the latter, it has been used quite broadly in research and consists the focus of the current study.

A typical metataxonomic analysis includes a process that combines laboratory and bioinformatic workflows. The steps involved in the laboratory process concern the collection of a microbiome sample, the DNA extraction, the library preparation based on the preferred rRNA gene marker and the massive sequencing with the NGS platform of choice. The bioinformatic workflow concerns the quality filtering of the resulted data, the clustering of sequences based on a specific clustering strategy and the taxonomic assignment to the representative sequence of each cluster.

There are a plethora of bioinformatic frameworks for the analysis of the microbiome data with Quantitative Insights Into Microbial Ecology (QIIME) being one of the most popular and thus, implemented in the current study (Caporaso et al., 2010; Bolyen et al., 2018). As a bioinformatic

framework, it contains a significant amount of algorithms and parameters to select and tweak, respectively, but studies such as Bokulich et al. (2013, 2018) have provided informative and useful benchmarks with the resulted balanced parameters being incorporated into QIIME as default parameters. Nevertheless, microbiome samples are subjects to different laboratory procedures and protocols and as such implementation of parameters must be evaluated. For that reason, a mock community, which represents a microbiome sample of known composition (Bokulich et al., 2016), consists a valuable tool in assessing both laboratory and bioinformatic workflows prior to establishment of parameters. There are many studies dedicated to mock communities, such as Yuan et al. (2012) where a mock community was used for the comparison of six common DNA extraction protocols, or Yeh et al. (2018) where mock communities were the tool for the establishment of a methodology that could verify similar performance between sequencing runs. However, the way that the current study differs from the rest is based on the fact that the main focus is given on assessing the effectiveness of QIIME balanced default parameters on our laboratory and bioinformatic workflows destined to the metataxonomic analysis of wine samples.

Wine samples are characterized by extremely dynamic microbial populations. During wine ageing, these populations tend to be quite sparse with most of the microorganisms being difficult to detect as they enter the viable but non-culturable state (VBNC) (Millet and Lonvaud-Funel, 2000), and thus making NGS technology the most appropriate detection tool. Therefore, sparse microbial communities are quite important since wine spoilage microorganisms may go undetected due to their low abundance and significantly alter the wine quality later on. For that reason, the mock community in the current study was chosen to be simple. Additionally to the main focus, the mock community will serve a double qualitative role on a series of aspects related to our workflows. Regarding the laboratory procedure, to evaluate 16S metataxonomic analysis on data produced by Ion Torrent and Illumina platforms, the impact of 18S and ITS amplicons on the metataxonomic classification and the effect of the PCR cycles during the library preparation on the downstream bioinformatic analysis of the Ion Torrent data. As far as the bioinformatic analysis is concerned, the mock community will assist in ascertaining the impact on classification of different quality filtering thresholds, the performance of different sequence clustering methods and the classification performance of two different algorithms. Moreover, we are examining the possibility of utilizing the confidence of the assigned taxonomy, as reported by the classification algorithms, as a tool for eliminating false positives.

2. METHODS

2.1. Laboratory Workflow

Two microbial community standards from ZymoBIOMICS^{TM} with the same microbial composition of 8 prokaryotes and 2 eukaryotes and impurity level $<$ 0.01% have been used. The first standard contained DNA extracted from pure cultures (DNA standard D6305 200 ng), whereas the second standard was constructed by pooling pure cultures (Microbial Community standard D6300). The microbial species along with the 16S theoretical relative abundance, as provided by the standards specifications, are given in **Table 1**. The theoretical relative abundances have been calculated by the standards provider taking into consideration differences in the number of copies each amplicon has among the species. However, such correction is rendered impossible when estimating relative abundances in real wine samples. Therefore, the estimated relative abundances have not been corrected in order to examine the amount of deviation between estimated and ideal relative abundance. The aim of using the DNA standard (DS) was to assess the performance of different PCR primers and amplicons used with the NGS platforms, the impact of PCR cycles on the number of chimeric sequences in the Ion Torrent platform, as well as the performance of the bioinformatic pipelines at reconstructing the 16S theoretical relative abundance as well as assigning correct taxonomy to the eukaryotic DNA. The additional goal of using the culture standard (CS) was to ascertain the effectiveness of the in-house DNA extraction protocol that follows the recommended procedure of the DNeasy Plant Mini kit (Qiagen, Hilden, Germany), including three bead-beating steps for 3 minutes in a FastPrep-24 bead beater (MP Bio, Solon, OH) (Lleixà et al., 2018).

TABLE 1 | Culture and DNA standard microbial composition of the mock communities used during the current study and 16S theoretical relative abundance.

Species	NRRL accession NO.	Theoretical composition of 16S rRNA(%)	
		Culture standard	DNA standard
Pseudomonas aeruginosa	B-3509	4.2	4.6
Escherichia coli	B-1109	10.1	10.0
Salmonella enterica	B-4212	10.4	11.3
Lactobacillus fermentum	B-1840	18.4	18.8
Enterococcus faecalis	B-537	9.9	10.4
Staphylococcus aureus	B-41012	15.5	13.3
Listeria monocytogenes	B-33116	14.1	15.9
Bacillus subtilis	B-354	17.4	15.7
Saccharomyces cerevisiae	Y-567	-	-
Cryptococcus neoformans	Y-2534	-	-

Based on ZymoBIOMICS^{TM}, the strain information was extracted from the website of the Agricultural Research Service Culture Collection and can be accessed with the NRRL accession number (NRRL, https://nrrl.ncaur.usda.gov/).

Amplicon based sequences were generated by two different platforms, Ion Torrent (Centre for Omics Sciences, Reus, Spain) and Illumina (Centre for Genomic regulation, Barcelona, Spain). In the case of Ion Torrent, the sequencing libraries were prepared in the in-house laboratory of the University Rovira i Virgili using both the DNA and culture standard. For the libraries creation, the 16S rRNA region was amplified by PCR with the primers 515F and 806R (Caporaso et al., 2011) whereas the 18S rRNA region was amplified using the primers FR1 and FF390 (Prevost-Boure et al., 2011). Since a positive correlation between PCR cycles and amount of chimeric sequences has been reported (Ahn et al., 2012), 30 and 45 PCR cycles were used for the libraries

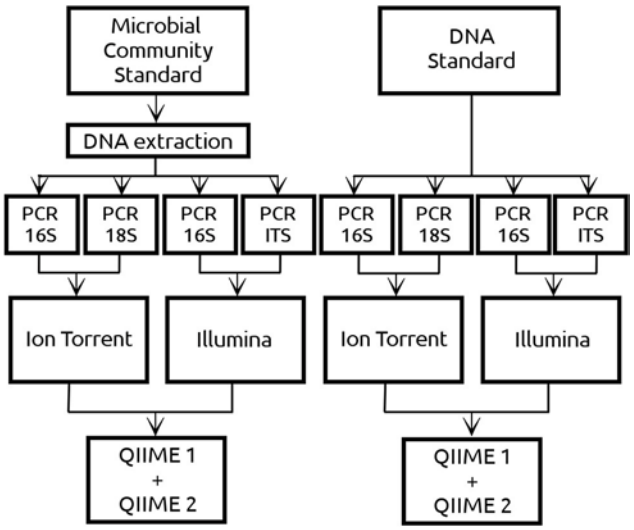


FIGURE 1 | Two commercial mock community standards from ZymoBIOMICS™ with exactly the same microbial composition of 8 prokaryotes and 2 eukaryotes have been used in the current study. The Microbial Community standard (referred as CS) consisted of microbial cells from which DNA was extracted using an in-house DNA extraction protocol. The DNA standard (referred as DS) contained DNA from the same 10 microbial cells as the CS but extracted by ZymoBIOMICS™. Both standards were sequenced using Ion Torrent and Illumina platforms. Regarding the DNA from the prokaryotic cells, both platforms sequenced the 16S amplicon. Regarding the DNA from the eukaryotic cells, Ion Torrent sequenced the 18S amplicon whereas Illumina the ITS amplicon. In the case of Ion Torrent 30 and 45 PCR cycles have been implemented in both amplicons, whereas in Illumina only 30 PCR cycles were implemented. Sequencing data derived from both NGS platforms have been analyzed using QIIME 1 and QIIME 2.

creation. The PCR products were purified using GeneRed Size selection Kit (Qiagen, Hilden, Germany) and sent to COS for sequencing with the 530 chip using the Gene Studio S5 System of the Ion Torrent platform. On the other side, the DNA standard and extracted DNA from the culture standard were sent directly to CRG to be sequenced by Illumina MiSeq 2x300 yielding paired end sequences for the v3 region of the 16S [primers 341F and 785R, Herlemann et al. (2011)] and for the ITS region [primers ITS1F/ITS2R, White et al. (1990)]. Schematic representation of the experimental design is given in **Figure 1**.

The Ion Torrent platform generated in average 300 bp reads for the 16S amplicon and 350 bp reads for the 18S amplicon, with an average Phred33 quality score of 29 and 27, respectively. On the other hand, Illumina generated in average 300 bp reads for both amplicons with an average Phred33 quality score of 36 for both 16S and ITS forward reads and 34 and 35 for the 16S and ITS reverse reads, respectively. Due to the fact that the Phred33 quality of the Ion Torrent reads dropped below 10 in positions located in the middle of the read, two filtering strategies were applied. One applying a quality threshold at 10 (Q10) and one at 20 (Q20). The motivation behind these two strategies was to examine whether higher number of sequences or higher overall quality will produce better results. Contrarily, for the Illumina reads, only the Q20 threshold was applied.

2.2. Bioinformatic Workflow

Bokulich et al. (2013) benchmarked different quality filtering strategies with QIIME 1 and Bokulich et al. (2018) benchmarked the performance of difference classification algorithms between QIIME 1 and QIIME 2. Therefore, the bioinformatic pipelines were based on two versions of QIIME, QIIME 1 (version 1.9.1) and QIIME 2 (version 2018.2), with the processing and taxonomic assignment steps mentioned in **Table 2**. Along with QIIME, bioinformatic tools such as FastQC (Andrews, 2010), Trimmomatic (Bolger et al., 2014) and FLASH (Magoč and Salzberg, 2011) were executed externally.

From the default parameters of QIIME 1 for the quality filtering of raws reads, only the Phred33 quality threshold was altered. Generally, the quality filtering concerned discarding reads with consecutive bases above a given Phred33 threshold but occupying <75% of the total read length, truncating reads at positions with more than 3 consecutive bases with Phred33 quality less than the desired and reassessing the discarding rule after truncation. Due to the fact that QIIME 1 quality filtering steps require the sequences to be multiplexed, for the demultiplexed Illumina sequences the quality filtering steps of QIIME 1 were replicated in Trimmomatic. Moreover, the DADA2 algorithm (Callahan et al., 2016), as incorporated into QIIME, truncated reads at the first base instance of undesired quality and discarded reads with >2 expected errors. An additional filtering step was implemented by removing chimeric sequences with VSEARCH UCHIME de novo (Rognes et al., 2016) or DADA2.

Regarding the Illumina reads two clustering methods were applied. One that creates clusters of sequences, called operational taxonomic units (OTU) based on a similarity threshold (Rideout et al., 2014) and one that defines sequence variants called amplicon sequence variants (ASV) (Callahan et al., 2017). The

TABLE 2 | Bioinformatic pipelines based on NGS platform and method of clustering used during this study for comparison of their performance over the mock community standards.

Ion Torrent OTU	Illumina OTU	Illumina ASV
Barcode extraction[a]	Paired ends merging[c]	Paired ends merging[c]
Quality filtering (Q10 or Q20)[a]	Quality filtering (Q20)[d]	DADA2 quality filtering (Q20)[b]
Reads dereplication[b]	Reads dereplication[b]	DADA2 reads dereplication[b]
Open reference OTU[b]	Open reference OTU[b]	DADA2 Chimeras filtering (only ITS)[b]
Chimeras filtering[b]	Chimeras filtering[b]	DADA2 ASV[b]
SKLEARN classifier training[b]	SKLEARN classifier training[b]	SKLEARN classifier training[b]
SKLEARN taxonomy assignment[b]	SKLEARN taxonomy assignment[b]	SKLEARN taxonomy assignment[b]
BLAST+ taxonomy assignment[b]	BLAST+ taxonomy assignment[b]	BLAST+ taxonomy assignment[b]

[a] *QIIME 1 (version 1.9.1).*
[b] *QIIME 2 (version 2018.2).*
[c] *FLASH.*
[d] *Trimmomatic.*

OTU method produces an OTU-table where, for each sample, the number of sequences in each OTU has been recorded (Rognes et al., 2016), whereas the ASV method is related with an ASV-table of the frequency that each ASV has been observed in each sample (Callahan et al., 2016). OTUs containing <10 sequences across all samples were filtered-out as noise (Giordano et al., 2018), and the similarity threshold for the OTU clustering was set to 99% as this threshold returns more comparable results between OTU and ASV (Van Der Pol et al., 2018).

For the metataxonomic classification the database SILVA (version 132) has been the source of taxonomy for the 16S and 18S amplicons (Quast et al., 2012) as it is the most recent and updated database, whereas the ITS taxonomy relied on the UNITE database (version 7.2) (Nilsson et al., 2018). The taxonomic assignment was carried out by two algorithms, the k-mer based multinomial naive Bayes algorithm integrated in the Python Scikit-learn library (SKLEARN) (Pedregosa et al., 2011) and the Basic Local Alignment Search Tool+ (BLAST+) algorithm which represents an enhanced version of the very popular BLAST algorithm available from 1997 (Camacho et al., 2009). Both algorithms report a confidence percentage, with the SKLEARN algorithm referring to the amount of confidence for the taxonomy assigned at a specific taxonomic level and BLAST+ referring to the fraction of top hits that matched the consensus taxonomy at a given level. As SKLEARN represents a machine learning approach, the additional flexibility provided was to assign taxonomy after training the algorithm with extracted reference sequences from the SILVA and UNITE databases using the aforementioned PCR primers and trimmed to a length equal to the maximum length of the reads after quality filtering. The training process of SKLEARN is based on k-mers where the value 7 was used as it is the default balanced QIIME 2 parameter. In relaxed terms, during the training process SKLEARN splits each reference sequence into a series of overlapping heptamers and assigns a level of taxonomy to a given collection of heptamers. Later on, during the classification process SKLEARN splits each sequence once again into a collection of overlapping heptamers, and tries to assign a level of taxonomy by taking into consideration the collections of heptamers from the reference sequences. The balanced default parameters of BLAST+ remained unaltered whereas the performance of SKLEARN improved after reducing the confidence parameter from the default 0.7 value down to 0.5.

3. RESULTS

Figure 2 shows the number of sequences for each sample after applying Phred33 quality filtering and removing chimeras. For the Ion Torrent a mild filtering was applied after setting the quality threshold at Q10 with an average of 8.6% of the sequences filtered, across all samples, for the 16S amplicon and 14.1% for the 18S whereas at Q20 an average of 62 and 72.4% was removed, respectively. An additional average of 13.5% of the sequences were identified as chimeras for the 16S amplicon and 1.2% for the 18S at Q10, while at Q20 the identified chimeras were 5.9 and 1.3%, respectively. Considering the PCR cycles, their impact on the production of chimeras was not clear for the 16S amplicon as at Q10, 45 cycles generated 3.5% more chimeras than 30 cycles for the CS but for the DS they produced 4.2% less. The same pattern repeated for the 16S amplicon at Q20 with 45 cycles of the CS producing 1.6% more chimeras but for the DS 3.5% more chimeras produced from 30 cycles. On the other hand, the difference was more apparent for the 18S amplicon producing more chimeras at 45 than 30 cycles, but the difference was marginal representing only 1.6% of the sequences in average (**Figure 2A**).

For the Illumina platform, the merging of the paired ends caused a $\approx$ 2% loss of reads for the 16S amplicon in both standards, whereas for the ITS amplicon of the DS the loss was 38%. Due to the fact that the sequencing of the ITS amplicon for the CS generated very low amount of sequences which had very low Phred33 quality, this sample was excluded from the study. This was the additional reason for not reporting the theoretical abundance of 18S and ITS amplicons, along with the fact that from the two standards only the CS reports 18S theoretical abundance in the specifications. However, research interest still remained on examining whether the classification algorithms could assign correct taxonomy to the eukaryotic DNA and which amplicon of the two improves classification performance. For the 16S amplicon of the CS, the Illumina OTU pipeline removed 1.2% of sequences during the quality filtering step and an additional 23.7% was identified as chimeras. The pipeline performed quite similar for the DS removing 1 and 17.9%, respectively. On the contrary, for the 16S amplicon of the two standards the Illumina ASV pipeline identified $\approx$ 80% of the sequences as chimeric. This high percentage could be justified in cases where non-biological nucleotides, such as primers or adapters, have not been removed prior to analysis [1], but since this rationale did not hold for the given dataset, the chimera filtering step was omitted for both standards. Therefore, the only loss was during the quality filtering with both standards losing $\approx$ 5% of sequences. Regarding the ITS amplicon of the DS, the Illumina OTU pipeline filtered 0.8% of sequences based on quality but did not identify any chimeras, and the Illumina ASV pipeline removed 1.9% during quality filtering and a further 5% during chimera filtering (**Figure 2B**).

The metataxonomic classification was performed at genus level since accurate classification at species level is a known limitation of rRNA amplicon sequencing due to the fact that it is a highly conserved region (Sentausa and Fournier, 2013). This limitation became apparent also in the current study as the only bacterium identified consistently and accurately at species level was *Listeria monocytogenes* whereas *Salmonella* was the only one whose classification never reached species level. From the rest, *Bacillus* demonstrated the highest variability with overall 7 different species being identified, 5 species for *Staphylococcus* and *Pseudomonas*, and $\leq$ 3 for *Escherichia*, *Lactobacillus*, and *Enterococcus*. Although this broad variability concerned the OTU clustering method, the variability in the ASV method was more constrained including only the cases of either correct species

[1] https://benjjneb.github.io/dada2/tutorial.html

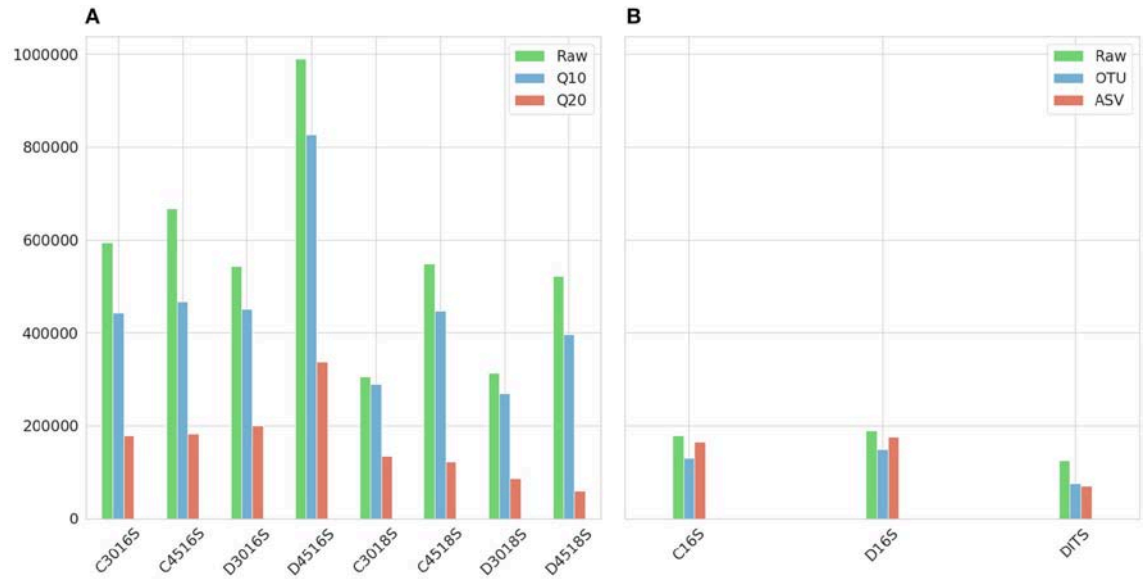


FIGURE 2 | Number of sequences resulted after applying quality and chimeras filtering. **(A)** Ion Torrent. First letter of the sample names (C or D) represents type of mock community standard (Culture or DNA). What follows is the number of PCR cycles (30 or 45) with the amplicon (16S or 18S) at the end. The raw number of sequences are represented in green and in red and blue the two Phred33 quality filtering strategies Q20 and Q10, respectively. **(B)** Illumina. First letter of the sample names (C or D) represents type of mock community standard (Culture or DNA) with the amplicon (16S or ITS) at the end. The raw number of sequences are represented in green and in red and blue the sequences resulted from the filtering steps of the Illumina ASV and Illumina OTU pipeline, respectively.

identification, no species identification or species identification as *uncultured bacterium*.

Figures 3–6 depict 16S estimated relative abundance (orange color) being juxtaposed against theoretical relative abundance (blue color) for both standards and NGS platforms. Overlapping between the two abundances is being represented with dark gray color and estimated abundance below 1% or undefined (0%) is being represented numerically. Excess of orange color at the bar edges denotes abundance overestimation whereas excess of blue color abundance underestimation. Next to each figure the taxonomic assignment confidence is being displayed as it has been reported by the classification algorithm at genus level (All). An additional step has been performed where the assigned taxonomies have been filtered by setting a confidence threshold which is displayed next to the unfiltered confidence. This threshold was initially set to 90% ($>$ 0.90) and gradually decreased until an optimal balance between amount of false positives and theoretical abundance reconstruction is achieved. Apart from **Figures 5B**, **6B,D** this confidence threshold matches the minimum unfiltered confidence reported by the classification algorithm giving an identical estimated relative abundance before and after confidence filtering as well as the same amount of false positives (FP).

For the Ion Torrent platform, SKLEARN failed to identify *Salmonella* regardless quality filtering threshold, PCR cycles or standard type, while achieved best performance with the DS, 45 PCR cycles, Q20 and confidence threshold 80% (**Figure 4G**). Overall, the maximum number of false positives was 2 with the genera *Carnobacterium*, *Citrobacter*, *Oenococcus*, and *Pediococcus* consisting the pool of false positives. At the same time, BLAST+ seems to have exhibited a better performance than SKLEARN with optimal performance also with the DS, 45 cycles and Q20 (**Figure 4H**), but generating higher amounts of false positives and requiring a lower confidence threshold for optimal performance. In general, BLAST+ proved to be more sensitive than SKLEARN with 5 as the maximum number of false positives and a persistent confidence threshold of 60%. The false positives identified by BLAST+ were the genera *Cedecea*, *Citrobacter*, *Enterobacter*, *Klebsiella*, *Oenococcus*, and *Pediococcus*.

With Illumina generated data, the landscape was more clear. Both pipelines, Illumina OTU and ASV, yielded similar results with both classification algorithms performing better with the DS (**Figure 6**). Once again BLAST+ held the best performance managing to approximate quite accurately the theoretical composition (**Figures 6B,D**). However, it demonstrated overall higher sensitivity producing more false positives with their number being affected by even a slight increase of the confidence threshold by just 1% from the minimum reported confidence of 69% (**Figures 5B**, **6B,D**). The pool of false positives for SKLEARN was comprising the genera *Acetobacter*, *Enterobacter*, and *Oenococcus*, whereas for BLAST+ the genera *Citrobacter*, *Acetobacter*, *Cronobacter*, *Enterobacter*, and *Oenococcus*. In general, although the relative abundance of the false positives remained below 0.01%, the only exemption was with the CS and the Illumina ASV pipeline where *Cronobacter* reached 0.3%. Moreover, even if the confidence level of the classification assignment was quite low for the false positives in both algorithms (60%), the genera that defied this trend were *Acetobacter*, *Enterobacter* and *Oenococcus* reaching as high as 90% confidence.

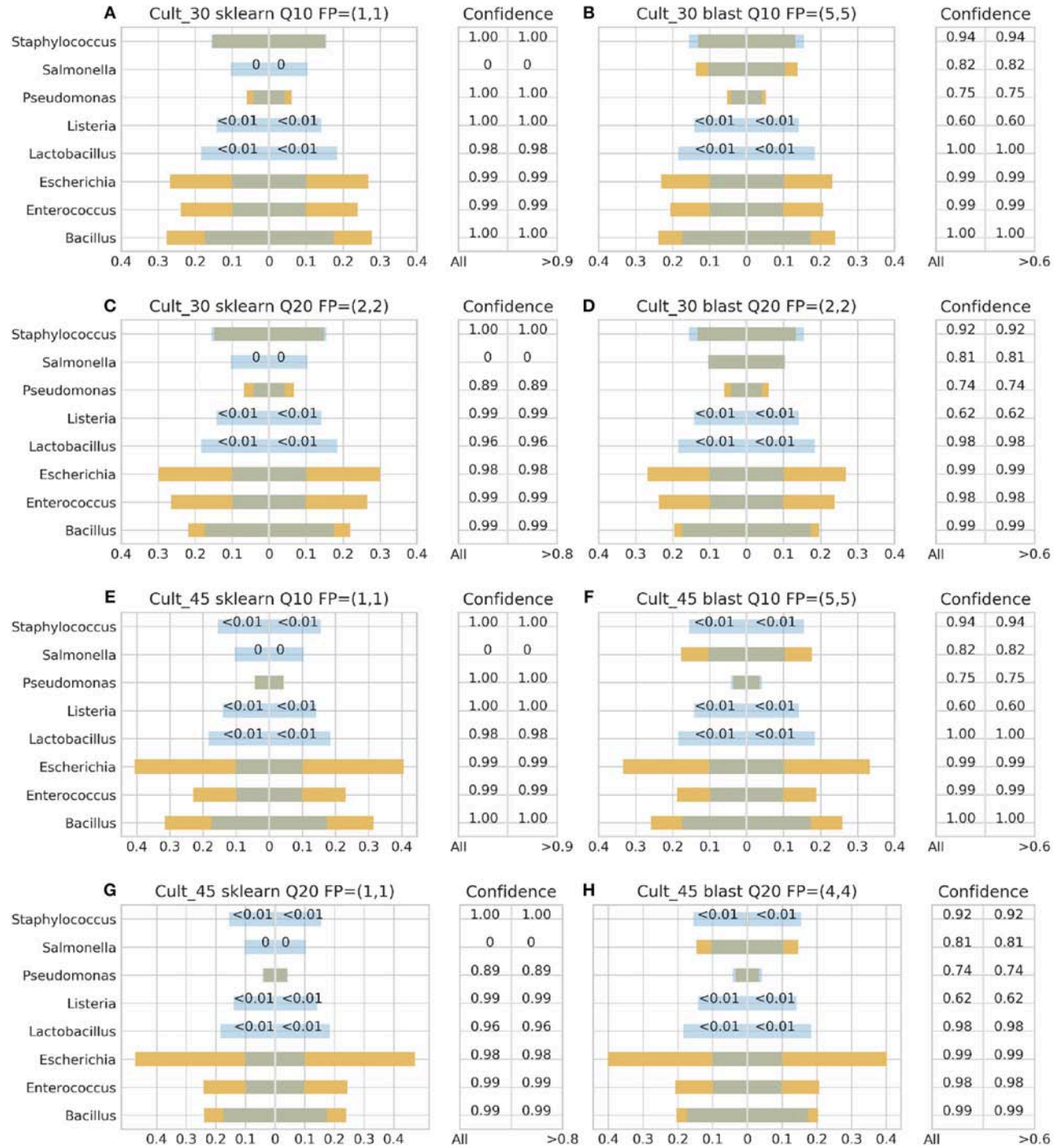


FIGURE 3 | 16S theoretical (blue color) and estimated (orange color) relative abundance for culture standard using Ion Torrent. Overlapping between the two abundances is being represented with dark gray color. Cult_30 and Cult_45 represent 30 and 45 PCR cycles, Q10, and Q20 Phred33 quality filtering threshold and FP false positives without (first number) and with confidence filtering (second number). Figures to the left **(A,C,E,G)** represent estimated abundance based on SKLEARN algorithm and to the right **(B,D,F,H)** based on BLAST+. Estimated relative abundance to the left side of 0 is based on unfiltered confidence (All) and to the right on filtered (> %).

With respect to fungi, none of the algorithms detected *Cryptococcus* regardless NGS platform or standard type, contrary to *Saccharomyces* which was detected though not always at species level. In both Illumina OTU and ASV pipelines, both algorithms exhibited similar performance by identifying only *Saccharomyces* with 100% confidence without yielding any false

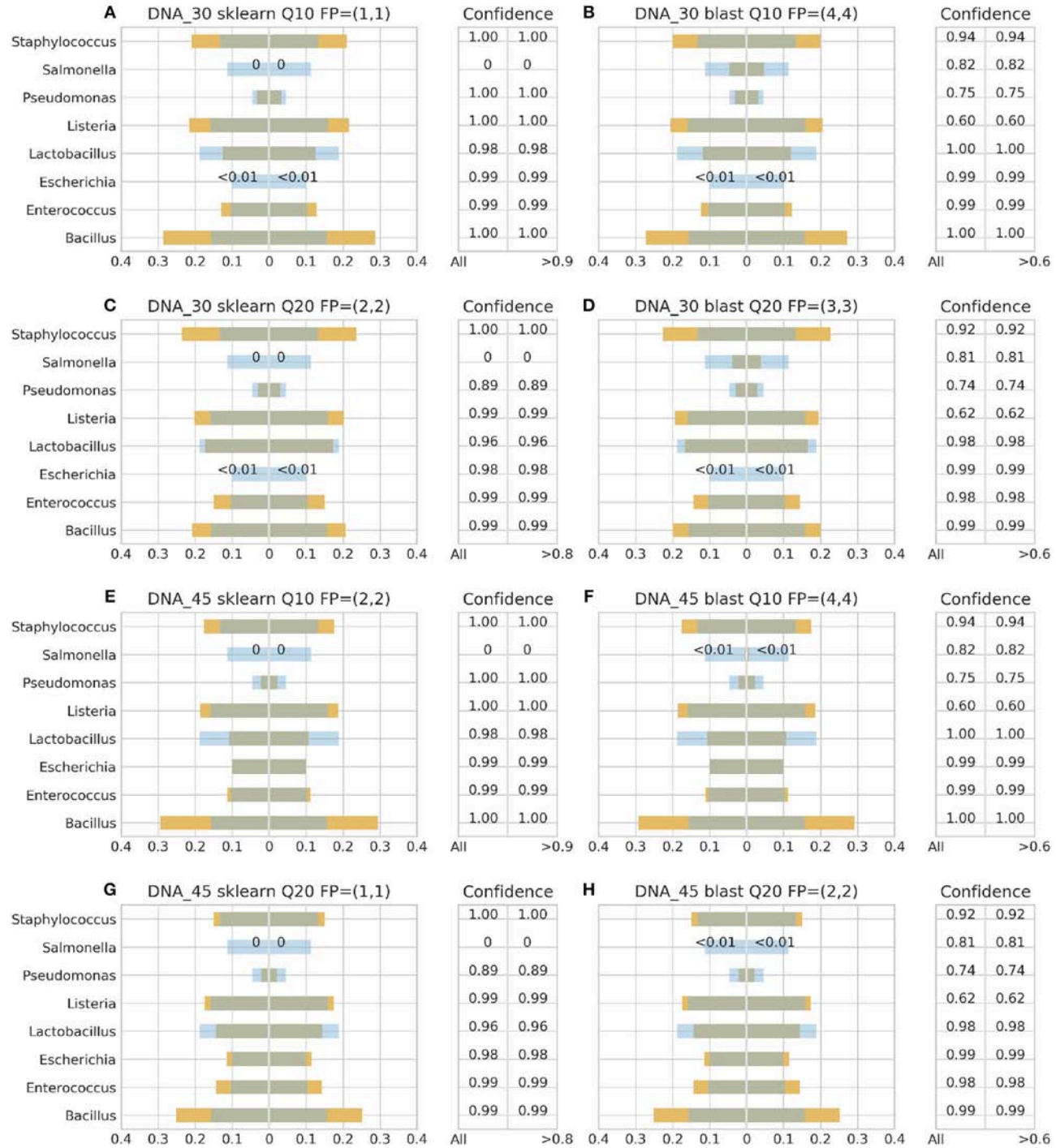


FIGURE 4 | 16S theoretical (blue color) and estimated (orange color) relative abundance for DNA standard using Ion Torrent. Overlapping between the two abundances is being represented with dark gray color. DNA_30 and DNA_45 represent 30 and 45 PCR cycles, Q10, and Q20 Phred33 quality filtering threshold and FP false positives without (first number) and with confidence filtering (second number). Figures to the left **(A,C,E,G)** represent estimated abundance based on SKLEARN algorithm and to the right **(B,D,F,H)** based on BLAST+. Estimated relative abundance to the left side of 0 is based on unfiltered confidence (All) and to the right on filtered (> %).

positives. On the other hand, BLAST+ in Ion Torrent managed to identify *Saccharomyces* with 99.9% confidence in both standards regardless quality threshold and PCR cycles, but produced *Zygosaccharomyces* as a false positive with CS at Q10 and 30 cycles and *Kazachstania* with DS at Q20 and 45 cycles having a 60% confidence in both cases. On the side of SKLEARN,

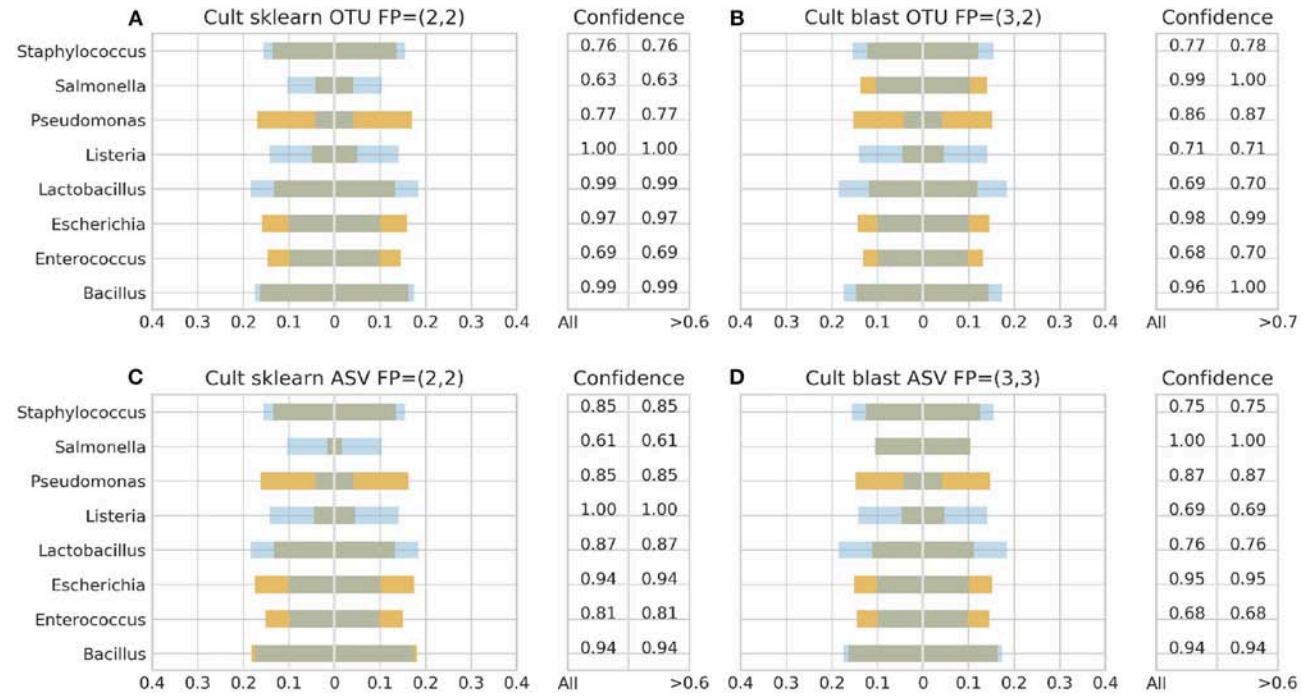


FIGURE 5 | 16S theoretical (blue color) and estimated (orange color) relative abundance for culture standard using Illumina. Overlapping between the two abundances is being represented with dark gray color. OTU and ASV represent Illumina OTU and Illumina ASV pipelines and FP false positives without (first number) and with confidence filtering (second number). Figures to the left **(A,C)** represent estimated abundance based on SKLEARN algorithm and to the right **(B,D)** based on BLAST+. Estimated relative abundance to the left side of 0 is based on unfiltered confidence (All) and to the right on filtered (> %).

Saccharomyces occupied ≈ 61% of the relative abundance in average across the different PCR cycles in both standards at Q10 with the rest of the abundance being occupied by a taxonomy assigned as *uncultured fungus*. At Q20, *Saccharomyces* occupied 99% of the relative abundance with the DS at 45 cycles and 50% in the rest of the samples, with the remaining abundance once again assigned as *uncultured fungus*. Although in the case of BLAST+ the false positives could be removed by raising the confidence threshold, in the case of SKLEARN confidence filtering did not improve the result as the confidence level was in average 90% for *Sacchraromyces* and 85% for the false positives.

4. DISCUSSION

A mock community represents a microbiome sample of known microbial composition and in the current study two types of mock community standards with the same species composition have become the tool for evaluating the effectiveness of QIIME balanced default parameters on metataxonomic analysis workflows destined to the analysis of wine aging samples. The evaluation was performed with QIIME framework and two classification algorithms, one representing a popular local alignment algorithm (BLAST+) and the other one a popular machine learning approach (SKLEARN). These two algorithms have been introduced for the first time in QIIME 2 and their performance compared to the classification algorithms of QIIME 1 have been benchmarked by Bokulich et al. (2018) where they exhibited similar as well as enhanced performance on different performance metrics. Moreover, Bokulich et al. (2013) in QIIME 1 benchmarked different quality-filtering strategies so as to provide guidelines for processing Illumina amplicon-based sequencing data. Although the suggested parameters of these studies have been incorporated as balanced default parameters in QIIME , microbiome samples undergo different laboratory procedures and protocols and thus these parameters should be evaluated prior to implementation. Therefore, the aim of the present study was to examine the effect of these parameters on a series of aspects related to our laboratory and bioinformatic workflows using a mock community and focusing on reconstructing the theoretical 16S relative abundance or yeast composition based on 18S and ITS amplicon sequencing. Furthermore, the mock community facilitated the qualitative assessment of other aspects such as the performance of the classification algorithms, the possibility of utilizing the reported taxonomic assignment confidence from the classification algorithms as a tool for eliminating false positives, the performance of Ion Torrent and Illumina NGS platforms with the 16S amplicon, the effect of PCR cycles on the analysis of Ion Torrent data, as well as the outcome of the in-house DNA extraction protocol by using a culture based standard (CS).

The 16S metataxonomic analysis of the CS approximated quite closely the outcome of the DS analysis in the Illumina platform, while it demonstrated an apparent variability in the

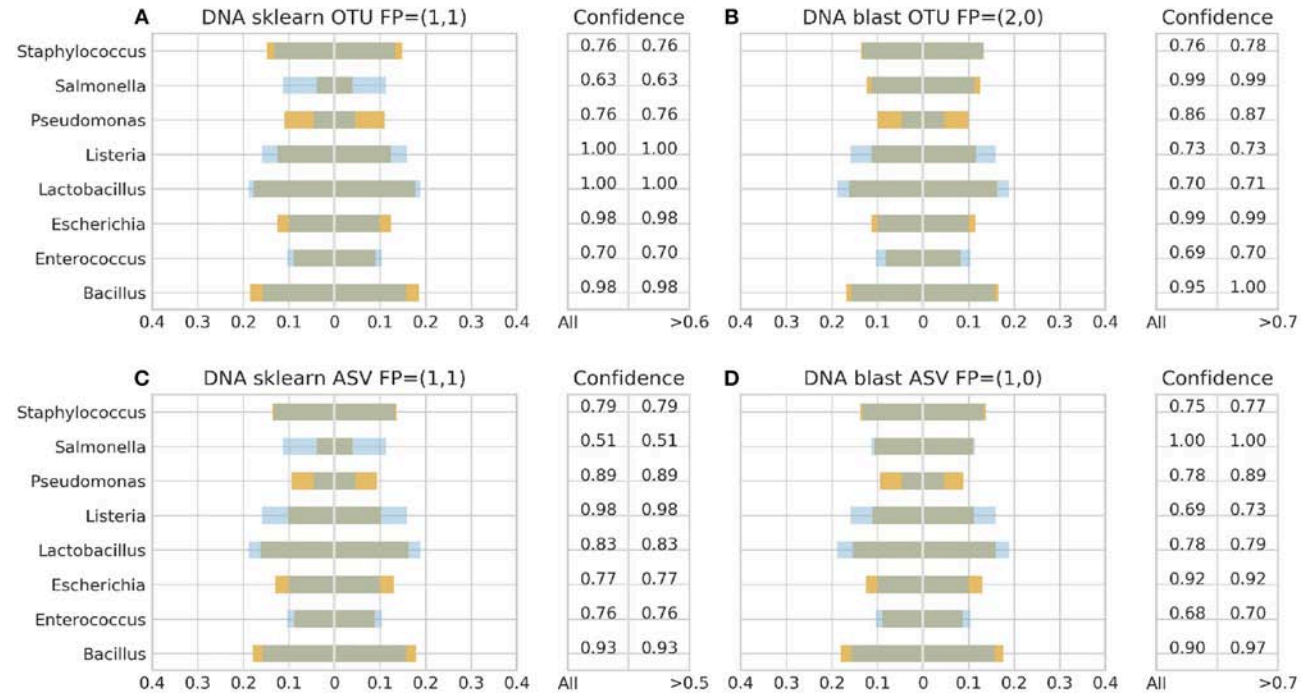


FIGURE 6 | 16S theoretical (blue color) and estimated (orange color) relative abundance for DNA standard using Illumina. Overlapping between the two abundances is being represented with dark gray color. OTU and ASV represent Illumina OTU and Illumina ASV pipelines and FP false positives without (first number) and with confidence filtering (second number). Figures to the left **(A,C)** represent estimated abundance based on SKLEARN algorithm and to the right **(B,D)** based on BLAST+. Estimated relative abundance to the left side of 0 is based on unfiltered confidence (All) and to the right on filtered (> %).

case of the Ion Torrent platform. On the other hand, the Ion Torrent 18S analysis produced similar results in both standards. This denotes that pinpointing a performance culprit among the NGS platforms, PCR protocols or bioinformatic pipelines is rendered difficult as a further variability is being added by the DNA extraction protocol. Regarding the discard of the ITS amplicon based sample of the CS due to low quality, it has been attributed to the poor performance of the DNA extraction protocol since good quality Illumina sequences were generated with the corresponding sample of the DS.

With Ion Torrent, both classification algorithms performed better with the DS linked to 45 PCR cycles and Q20 as a quality threshold signifying that optimal performance is more related to better overall sequence quality rather than higher amount of sequences as produced by the Q10 threshold. This could be associated with the fact that Q20 is related to 1% base call error rate while Q10 to 10% (Ewing and Green, 1998), indicating that low Phred33 quality threshold might lead to higher possibility of misclassification. Nevertheless, this result could not be easily attributed to the PCR cycles as 45 cycles in DS produced the highest amount of sequences among all samples and on the other hand in CS both algorithms favored 30 cycles. Moreover, the impact of PCR cycles on the amount of chimeric sequences was either marginal or unclear, however a negative correlation between quality threshold and amount of chimeras became apparent with the 16S amplicon, with fewer chimeras being identified at Q20 threshold. This indicates that a small increase of the PCR cycles does not influence greatly the production of chimeras and many of those chimeric sequences had overall low quality as they represent PCR artifacts. Similarly, slight difference on the production of chimeric sequences was also observed by a small increase of PCR cycles in the study of Ahn et al. (2012) when 25 PCR cycles were compared to 30 cycles, however great disparity on the amount of chimeras was observed between 15 and 30 cycles with the authors suggesting the lowest PCR cycles possible.

As Van Der Pol et al. (2018) suggested, setting the similarity threshold to 99% for the OTU clustering method produced similar results as the ASV method in Illumina, however the latter demonstrated a narrower variability of taxonomic assignment at species level. Furthermore, the omitted chimera filtering step in Illumina ASV pipeline for the 16S amplicon highlighted its importance as false positives above the impurity level of 0.01% were emerged. Additionally, the two NGS platforms presented different filtering behaviors at Q20 with Ion Torrent removing more sequences during the Phred33 quality filtering and less during chimera filtering, whereas Illumina performed the opposite. That could indicate that more chimeric sequences with high Phred33 quality score were generated with Illumina.

As a whole, BLAST+ exhibited better and more balanced performance in both NGS platforms than SKLEARN, however it demonstrated higher sensitivity producing more false positives and overall lower confidence regarding taxonomic assignment. The low amount of false positives generated by SKLEARN with the 16S amplicon could be associated with its training process as higher amount of reference sequences were extracted from the database with the PCR primers of this amplicon compared to 18S and ITS. Nonetheless, its enhanced performance with the Illumina data could be connected to the fact that its default parameters were linked with this NGS platform in the study of Bokulich et al. (2018). Moreover, the lack of false positives from both algorithms with the ITS amplicon could be explained by its higher specificity compared to 18S (Trtkova and Raclavsky, 2006), and overall the reported taxonomic assignment confidence from the algorithms could not lead to an effective filtering tool of false positives as some of the false taxonomies have been assigned with high confidence level.

5. CONCLUSIONS

Overall, the mock community standards have been proven a useful tool demonstrating good performance of QIIME balanced default parameters on our workflows especially with the Illumina platform. Nevertheless, the performance of the NGS platforms or the classification algorithms should not be considered deterministic since an exhaustive benchmarking process is needed for that purpose. As underlined by Bokulich et al. (2018), further fine-tuning of the QIIME default parameters with limited number of mock communities could lead closer to an overfitted rather than generalized performance. Moreover, a series of qualitative performance expectations could be proposed that could be summarized as better metataxonomic outcome when setting the Phred33 quality filtering threshold as high as possible, marginal difference in chimeras production between 30 and 45 PCR cycles, less false positives with ITS amplicon sequencing compared to 18S, similar performance between ASV and OTU clustering method when the clustering similarity threshold of the latter is set to 99% and more comparable results between Ion Torrent and Illumina platforms using the BLAST+ classification algorithm.

DATA AVAILABILITY

All raw sequencing data used in the current study have been deposited into Sequence Read Archive (SRA) repository under the BioProject accession number PRJNA524645. The raw data are publicly available from the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

AUTHOR CONTRIBUTIONS

AM and MP contributed to the experimental design, funding of the study and writing of the discussion section of the paper. DK performed the DNA extraction, bioinformatic analysis, and writing of the paper. However, all authors had a substantial, direct, and equal intellectual contribution to this study.

FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 713679 and from the University Rovira i Virgili (URV). Additional support has been received by the project AGL2015-73273-JIN of the Spanish Government.

REFERENCES

Ahn, J.-H., Kim, B.-Y., Song, J., and Weon, H.-Y. (2012). Effects of pcr cycle number and dna polymerase type on the 16s rrna gene pyrosequencing analysis of bacterial communities. *J. Microbiol.* 50, 1071–1074. doi: 10.1007/s12275-012-2642-z

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2 q2-feature-classifier plugin. *Microbiome* 6:90. doi: 10.1186/s40168-018-0470-z

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., et al. (2016). mockrobiota: a public resource for microbiome bioinformatics benchmarking. *MSystems* 1, e00062–16. doi: 10.1128/mSystems.00062-16

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nat. Methods* 10:57. doi: 10.1038/nmeth.2276

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., Al-Ghalith, G. A., et al. (2018). *Qiime 2: Reproducible, Interactive, Scalable, and Extensible microbiome Data Science.* Technical report, PeerJ Preprints.

Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform* 13, 669–681. doi: 10.1093/bib/bbx120

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 11:2639. doi: 10.1038/ismej.2017.119

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). Blast+: architecture and applications. *BMC Bioinform.* 10:421. doi: 10.1186/1471-2105-10-421

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. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A* 108(Suppl. 1), 4516–4522. doi: 10.1073/pnas.1000080107

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res* 8, 186–194.

Giordano, C., Spennati, F., Mori, G., Munz, G., and Vannini, C. (2018). The microbial community in a moving bed biotrickling filter operated to remove hydrogen sulfide from gas streams. *Syst. App. Microbiol.* 41, 399-407. doi: 10.1016/j.syapm.2018.04.001

Herlemann, D. P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., and Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the baltic sea. *ISME J.* 5:1571–1579. doi: 10.1038/ismej.2011.41

Lleixà, J., Kioroglou, D., Mas, A., and del Carmen Portillo, M. (2018). Microbiome dynamics during spontaneous fermentations of sound grapes in comparison with sour rot and botrytis infected grapes. *Int. J. Food Microbiol.* 281, 36–46. doi: 10.1016/j.ijfoodmicro.2018.05.016

Magoč, 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

Millet, V., and Lonvaud-Funel, A. (2000). The viable but non-culturable state of wine micro-organisms during storage. *Lett. Appl. Microbiol.* 30, 136–141. doi: 10.1046/j.1472-765x.2000.00684.x

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., et al. (2018). The unite database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.* 47, D259–D264. doi: 10.1093/nar/gky1022

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://www.jmlr.org/papers/v12/pedregosa11a.html

Prevost-Boure, N. C., Christen, R., Dequiedt, S., Mougel, C., Lelievre, M., Jolivet, C., et al. (2011). Validation and application of a pcr primer set to quantify fungal communities in the soil environment by real-time quantitative pcr. *PLoS ONE* 6:e24166. doi: 10.1371/journal.pone.0024166

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. doi: 10.1093/nar/gks1219

Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., et al. (2014). Subsampled open-reference clustering creates consistent, comprehensive otu definitions and scales to billions of sequences. *PeerJ* 2:e545. doi: 10.7717/peerj.545

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

Sentausa, E., and Fournier, P.-E. (2013). Advantages and limitations of genomics in prokaryotic taxonomy. *Clin. Microbiol. Infect.* 19, 790–795. doi: 10.1111/1469-0691.12181

Trtkova, J., and Raclavsky, V. (2006). Molecular-genetic approaches to identification and typing of pathogenic candida yeasts. *Biomed. Pap. Med. Fac. Univ. Palacky Olomouc Czech Repub.* 150, 51–61. doi: 10.5507/bp.2006.005

Van Der Pol, W. J., Kumar, R., Morrow, C. D., Blanchard, E. E., Taylor, C. M., Martin, D. H., et al. (2018). In silico and experimental evaluation of primer sets for species-level resolution of the vaginal microbiota using 16s ribosomal rna gene sequencing. *J. Infect. Dis.* 219, 305–314. doi: 10.1093/infdis/jiy508

White, T. J., Bruns, T., Lee, S., Taylor, J., et al. (1990). Amplification and direct sequencing of fungal ribosomal rna genes for phylogenetics. *PCR Prot. Guide Methods Appl.* 18, 315–322. doi: 10.1016/B978-0-12-372180-8.50042-1

Yeh, Y.-C., Needham, D. M., Sieradzki, E. T., and Fuhrman, J. A. (2018). Taxon disappearance from microbiome analysis reinforces the value of mock communities as a standard in every sequencing run. *MSystems* 3, e00023–18. doi: 10.1128/mSystems.00023-18

Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z., and Forney, L. J. (2012). Evaluation of methods for the extraction and purification of dna from the human microbiome. *PLoS ONE* 7:e33865. doi: 10.1371/journal.pone.0033865

Conflict of Interest Statement: 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.

Copyright © 2019 Kioroglou, Mas and Portillo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 04 June 2019
doi: 10.3389/fmicb.2019.01205

Metabolic Dependencies Underlie Interaction Patterns of Gut Microbiota During Enteropathogenesis

*Die Dai[1], Teng Wang[1], Sicheng Wu[1], Na L. Gao[1] and Wei-Hua Chen[1,2,3]**

[1] Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-Imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China, [2] College of Life Science, Henan Normal University, Xinxiang, China, [3] Huazhong University of Science and Technology Ezhou Industrial Technology Research Institute, Ezhou, China

In recent decades, increasing evidence has strongly suggested that gut microbiota play an important role in many intestinal diseases including inflammatory bowel disease (IBD) and colorectal cancer (CRC). The composition of gut microbiota is thought to be largely shaped by interspecies competition for available resources and also by cooperative interactions. However, to what extent the changes could be attributed to external factors such as diet of choice and internal factors including mutual relationships among gut microbiota, respectively, are yet to be elucidated. Due to the advances of high-throughput sequencing technologies, flood of (meta)-genome sequence information and high-throughput biological data are available for gut microbiota and their association with intestinal diseases, making it easier to gain understanding of microbial physiology at the systems level. In addition, the newly developed genome-scale metabolic models that cover significant proportion of known gut microbes enable researchers to analyze and simulate the system-level metabolic response in response to different stimuli in the gut, providing deeper biological insights. Using metabolic interaction network based on pair-wise metabolic dependencies, we found the same interaction pattern in two IBD datasets and one CRC datasets. We report here for the first time that the growth of significantly enriched bacteria in IBD and CRC patients could be boosted by other bacteria including other significantly increased ones. Conversely, the growth of probiotics could be strongly inhibited from other species, including other probiotics. Therefore, it is very important to take the mutual interaction of probiotics into consideration when developing probiotics or "microbial based therapies." Together, our metabolic interaction network analysis can predict majority of the changes in terms of the changed directions in the gut microbiota during enteropathogenesis. Our results thus revealed unappreciated interaction patterns between species could underlie alterations in gut microbiota during enteropathogenesis, and between probiotics and other microbes. Our methods provided a new framework for studying interactions in gut microbiome and their roles in health and disease.

Keywords: bacterial interaction patterns, metabolic interaction network, gut microbiota community, intestinal microbial ecology, enteropathogenesis, probiotics

OPEN ACCESS

Edited by:
Qi Zhao,
Shenyang Aerospace University, China

Reviewed by:
Josef Neu,
University of Florida, United States
Ravinder Nagpal,
Wake Forest School of Medicine, United States

***Correspondence:**
Wei-Hua Chen
weihuachen@hust.edu.cn

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

Received: *07 December 2018*
Accepted: *13 May 2019*
Published: *04 June 2019*

Citation:
Dai D, Wang T, Wu S, Gao NL and Chen W-H (2019) Metabolic Dependencies Underlie Interaction Patterns of Gut Microbiota During Enteropathogenesis. Front. Microbiol. 10:1205. doi: 10.3389/fmicb.2019.01205

INTRODUCTION

In recent decades, increasing evidence has strongly suggested that gut bacteria play an important role in human health and disease (Selber-Hnatiw et al., 2017; Jackson et al., 2018). Gut bacteria has been considered as a real tissue with its specific functions such as modulating the metabolic phenotype, influencing innate immunity, protecting against pathogens, and so on (Eckburg et al., 2005; Tomasello et al., 2017). Changes in the composition of the gut microbiota have been proven to be associated with many diseases (Jackson et al., 2018) including inflammatory bowel disease (IBD; Joossens et al., 2011; Matsuoka and Kanai, 2015; Chu et al., 2016; Sartor and Wu, 2017; Zuo and Ng, 2018), type 2 diabetes (Delzenne et al., 2015), obesity (Moreno-Indias et al., 2014; Tai et al., 2015), atherosclerosis (Drosos et al., 2015; Yamashita et al., 2015) and colorectal cancer (CRC; Aarnoutse et al., 2017; Liang et al., 2017; Russo et al., 2018). Among which, IBD (Miyoshi and Chang, 2017; Sartor and Wu, 2017), including both Crohn's Disease (CD) and ulcerative colitis (UC), is one of the most-studied imbalances between intestinal microflora and the immune system. Over the past 50 years, there was a dramatic increase in IBD (Sartor and Wu, 2017). In addition, patients with IBD are at increased risk of CRC, accounting for less than 2% of colon cancer cases yearly (Tilg et al., 2018). CRC, one of the most common cancers with the highest mortality worldwide, has also been reported to be associated with intestinal microflora (Zeller et al., 2014).

Gut microbes live as a community, sharing the common intestinal environment (Shetty et al., 2017). They interact with each other, maintaining the intestinal microbial flora in a state of equilibrium (Sommer et al., 2017). The composition of gut microbiota is thought to be largely shaped by interspecies competition for available resources along with cooperative interactions (Zelezniak et al., 2015). Diet is considered as one of the main drivers (De Filippo et al., 2010), with certain contributions from intrinsic metabolic dependencies. However, to what extent the changes could be attributed to external factors like diet of choice and internal factors such as mutual relationships among gut microbiota, respectively, are yet to be elucidated. Furthermore, it is still unclear how such intrinsic dependencies could contribute to the parthenogenesis of intestinal diseases such as IBD and CRC.

In this study, we performed systematic network analysis based on pairwise interspecies metabolic dependencies among gut microbes in IBD and CRC patients and compared that of the healthy controls. Network analysis has proven to be a valuable tool in exploring interactions between a set of items (nodes, such as individuals in a school, species in a complex food web, proteins in metabolic pathways) by biologists and scientists in other fields (Kim and Hastak, 2018), and has recently been applied to explore and identify microbial patterns that are generally difficult to detect in complex systems (Chow et al., 2014; Cardinale et al., 2015; Kong et al., 2018). Due to the advances of high-throughput sequencing technologies, flood of (meta)-genome sequence information and high-throughput biological data are available for gut microbiota and their association with intestinal diseases, making it easier to gain understanding of microbial physiology at the systems level (Covert et al., 2004). In addition, the newly developed genome-scale metabolic models that cover significant proportion of known gut microbes enable researchers to analyze and simulate the system-level metabolic response in response to different stimuli in the gut, providing deeper biological insights (Zhang and Hua, 2016; Magnusdottir et al., 2017; van der Ark et al., 2017). Based on these data, we revealed unappreciated patterns in gut microbes of IBD and CRC patients and healthy controls, and were able to accurately predict the majority of the changes (i.e., decreased or increased) in the gut microbiota during enteropathogenesis. As compared with co-occurrence network (Cardinale et al., 2015), which has been widely applied in the identification and characterization of interspecies interactions among gut microbes, our metabolic dependency network is a directional network and can provide more information with considering the interaction between the bacteria. We thus concluded that metabolic dependencies underlie interaction patterns of gut microbiota community during enteropathogenesis, and believed that our methods could provide a new framework for studying interactions in gut microbiome and their roles in health and disease.

MATERIALS AND METHODS

Data Collection

Pair-Wise Interactions (Metabolic Dependencies) of Human Gut Microbes

Genome-wide metabolic models for 773 human gut microbes were obtained from Stefanía et al. (Magnusdottir et al., 2017). Pairwise interactions, i.e., changes *in silico* growth rates of two co-culturing microbes as compared with that of cultured alone were calculated using the methods described in the literature (Magnusdottir et al., 2017).

Briefly, genome-scale metabolic models of 773 human gut microbes described in literature (Magnusdottir et al., 2017) were reconstructed based on comparative genomics and enrichment literature-derived experimental data. Through a combination of detailed biochemical information from genome annotations and literature resources, genome-scale metabolic models can be constructed. The gene-protein-reaction (GPR) relationships are annotated in the metabolic modes with mass- and energy-balanced reactions. Furthermore, other omics data such as transcriptomic and proteomic data could be integrated into the model, making the model more informative. Additionally, pairwise simulations were performed on every pair of 773 microbes (298,378 pairs). Single and pairwise *in silico* growth rates were calculated on two different diets (Western and High fiber diet).

Based on these growth rates, we calculated the "weight" of the interaction between bacteria using the following equation, $w = \mathrm{Log}_2 \frac{P}{S}$, where P stands for growth rate of the species of interest when co-cultivated with another bacterium (paired growth rate) and S stands for growth rate when cultivated alone. A "w" value of 0 indicates the growth rate of a bacterium is

not changed by the other co-cultivated bacterium; a positive (negative) value of "*w*" indicates the growth rate can be promoted (inhibited) by the co-cultivated bacterium. The interactions between two bacteria are thus bi-directional.

Gut Metagenomic Data of IBD and CRC Patients and Healthy Controls

In total three metagenomic datasets, including two for IBD and one for CRC, were obtained from the European Nucleotide Archive (ENA; Leinonen et al., 2011)[1] database.

The first IBD datasets (referred to as IBD1 in our study) are available from ENA under the accession of ERP005534. It contained ten IBD and ten healthy individuals whose fecal microbiome compositions were determined using Illumina HiSeq 2500.

The second IBD datasets [ENA accession ID: SRP002423; referred to as IBD2 (NIH HMP Working Group et al., 2009; Noecker et al., 2016) in our study] contained 14 healthy samples and 20 disease samples; their fecal samples were sequenced using a 454 GS FLX Titanium pyrosequencer. This study is a part of the NIH Human Microbiome Project (HMP).

The third CRC datasets [ENA accession ID: ERP005534; referred to as CRC (Zeller et al., 2014) in our study] contained fecal samples of 53 patients and 61 healthy controls. In this study, metagenomic sequencing of fecal samples was used to identify potential markers for distinguishing CRC patients from tumor-free controls. The detailed description about the experiments actually entailed can be found in the literature (Zeller et al., 2014). In brief, fresh stool samples were collected and genomic DNA was extracted using the GNOME DNA Isolation Kit (MP Biomedicals). Then library preparation for metagenomic sequencing was automated and adapted on a Biomek FXp Dual Hybrid. And metagenomic sequencing was performed on the Illumina HiSeq 2000/2500 platform.

Read Processing and Quality Control

Trimmomatic (Bolger et al., 2014) was used to remove adaptors and low quality bases (trimming) from the Illumina paired-end and single-end reads. For Roche/454 sequence data, QTrim (Shrestha et al., 2014) was used for trimming. FastQC (Andrews, 2014)[2] was then used for quality control prior to downstream analysis; the generated HTML report files were manually examined for possible problems in the raw and processed data. The usable trimmed data were referred to as "Clean Data," and were used for downstream analysis.

Species Identification and Composition Analysis of Metagenomic Data

MetaPhlAn2 (Metagenomic phylogenetic analysis version 2; Truong et al., 2015) was used for the taxonomic composition analysis on the Clean Data with default parameters. MetaPhlAn2 can efficiently profile the composition of microbial communities with species level resolution.

Differential Abundance Analysis Between Disease and Healthy Samples

Wilcoxon Rank Sum test was used to identify differentially abundant species between patients and healthy controls. The detailed results are available in **Supplementary Table S1**. **Supplementary Figure S1** shown in is the boxplot of the relative abundances of identified species in IBD1 patients (red) and healthy controls (blue); the red (blue) dots under the box plots represent a significant decrease (increase) in the abundance in disease group. The classification of the bacteria (Commensal, Pathogen, and Probiotic) is provided by the literature (Magnusdottir et al., 2017), which is shown in **Supplementary Table S2**.

Construction and Characterization of Metabolic Dependency Network for Disease and Healthy Controls

The metabolic dependency networks were constructed using pairwise interactions and consisted of nodes and edges. Networks were constructed for each of the three datasets we collected, and separately for patients and healthy controls. For each network, the nodes were microbial species selected from the union of the top 50 most abundance species in patients and the respective healthy controls, whose combined account for more than 90% of the total abundances of all species, while the edges were pairwise interactions ("weights") between two connected species. To account for the impact of diets [Western and High fiber diet, as described in the literature (Magnusdottir et al., 2017)], two networks were constructed for each of the patient and control groups. At the end, four networks were obtained for each dataset. An open-source tool, Gephi (Bastian et al., 2009), was used for network visualization and analysis.

Statistics

All statistical analysis and plots were performed in R version 3.4.3[3]. Mann–Whitney and Chi square test were used to analyze differences between groups. The p-value < 0.05 was considered significant.

RESULTS

Construction of Metabolic Dependency Network for Gut Microbiota During Enteropathogenesis

The flow chart of the methods used is shown in **Supplementary Figure S2**. We collected gut metagenomics data from in total three published datasets, including two for IBD (NIH HMP Working Group et al., 2009; Noecker et al., 2016) and one for CRC (Zeller et al., 2014), each with different numbers of patients and healthy controls (see section "Materials and Methods" for details). We first constructed a metabolic dependency network for each of the sample groups (i.e., patients and controls). Briefly,

[1] http://www.ebi.ac.uk/ena

[2] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3] www.r-project.org

MetaPhlAn2 was used for the taxonomic composition analysis on the clean data with default parameters. The nodes in the network are microbial species selected from the union of the top 50 most abundance species that together account for more than 90% of the total abundances of all species in healthy and disease groups, while the edges represent pairwise interactions between two connecting species. The weight of the edge is the absolute value of the influence, which equals to log2-transformed growth rates change between co-culturing and single-growth (i.e., the growth rate when cultivated alone). The edges are thus directional; depending on the thresholds of the weights of the edges, there could be two edges connecting two neighboring nodes in the network, with each representing the impact of co-culturing as compared with the respective single-growth. Because the growth rate under co-cultured conditions could be slower than that of the single-culturing, we used red (green) to represent the increased (decreased) growth rate under co-cultured conditions.

For each of the three studies from where our data were obtained, we constructed networks for the patients and the respective healthy controls separately. To account for the impact of diet [western diet and a high fiber diet, as described in Magnusdottir et al. (2017)], we constructed two networks each of the patient and control groups. At the end, we obtained four networks for each dataset. In this study, we described the results of IBD1 in western diet as an example, other datasets produced approximately the same results which were not shown here.

Network Centrality Analysis Revealed Probiotics Are Among the Top Important Nodes

As shown in **Figure 1**, we included dependencies of weight greater than four to build the network for healthy and disease groups respectively. We used gray, green, red, and pink to indicate nodes for commensal, probiotic, pathogenic, and opportunistic pathogenic bacteria, respectively, using classifications from a public dataset (Magnusdottir et al., 2017). To identify subclusters in which nodes are more densely connected than to the rest of the network, we used a modularity algorithm (a "community" detection technique) implemented in Gephi (Bastian et al., 2009) and identified two main subclusters (**Figure 1**); among which, one was mainly composed of probiotics bacteria, while the other was mostly composed of species of the genus of bacteroides. Surprisingly, we found that some pathogen bacteria, such as some strains of Escherichia coli, was also included in the probiotics subcluster and had a notable inhibitory effect on the probiotics included (**Figure 1**).

We then checked the top important nodes in the metabolic dependency network. We used the Gephi's PageRank algorithm (Chen et al., 2007) to rank the nodes. In addition to network centrality, PageRank also considers both the inbound and outbound links, which is suitable for analyzing our metabolic dependency network. Strikingly, we found that most of the top 20 bacteria were probiotics (10/11 in health and 11/11 in disease states), as shown in **Supplementary Tables S3**, **S4** and **Figure 2**. These results thus indicate that probiotics may play important roles in the metabolic dependency network.

Growth of Probiotics Was Strongly Inhibited by Other Bacteria in Both Patients and Healthy Controls

Strikingly, we found that the growth of probiotics topped the centrality analysis was strongly inhibited by themselves and others; we found similar results in patients and healthy controls. As shown in **Figure 3**, we divided the interactions into four groups. First, the background group includes interactions among bacteria excluding the probiotics. Second, the within group includes interactions among probiotics. Third, the affecting group includes the impacts of probiotics to other bacteria. Fourth, the affected group includes the impacts of other bacteria on probiotics. We found that the weight scores were significantly lower in the "within" and "affected" groups as compared with the other two; we found similar trends in both patients and the controls (**Figures 3B,D**, respectively; Wilcoxon Rank Sum test). Similarly, we found that the proportion of inhibitory effects in the "affected" were significantly higher than other three groups (**Figures 3A,C**; Chi-square test). The "with" group contained significantly higher proportion of inhibitory effects than the "affecting" group; its proportion was also higher than that of the background, although the difference was not significant. These results indicate that although probiotics are mostly beneficial to the host, they often face competition from other probiotics and are clearly not welcomed by other.

Disease-Enriched Bacteria Are Boosted by Themselves as Well as Other Bacteria

We found that the growth of bacteria whose abundances were significantly increased in patients (then were hence referred as to "disease-enriched bacteria") could be promoted by themselves as well as by others. We divided pairwise interactions into four groups. First, the background group contains interactions excluding the disease-enriched bacteria and the probiotics. Second, the within-group includes interactions among the disease-enriched bacteria. Third, the affecting group includes the impacts of disease-enriched bacterial on others. Fourth, the affected group includes the impact of other bacteria on the disease-enriched ones. As shown in **Figure 4**, there were significantly more promoting affects in the second and third as compared with other two groups (**Figures 4B,C**), indicating a marked difference of the disease-enriched bacteria as a group as compared to others.

TABLE 1 | The recognition accuracy for the three datasets (analyzed in different diet).

	Accuracy (%)
IBD10_mphlan_HF	53.10
IBD10_mphlan_W	65.60
twin_mphlan_HF	72.70
twin_mphlan_W	63.60
CRC_mphlan_HF	57.14
CRC_ mphlan_W	47.62

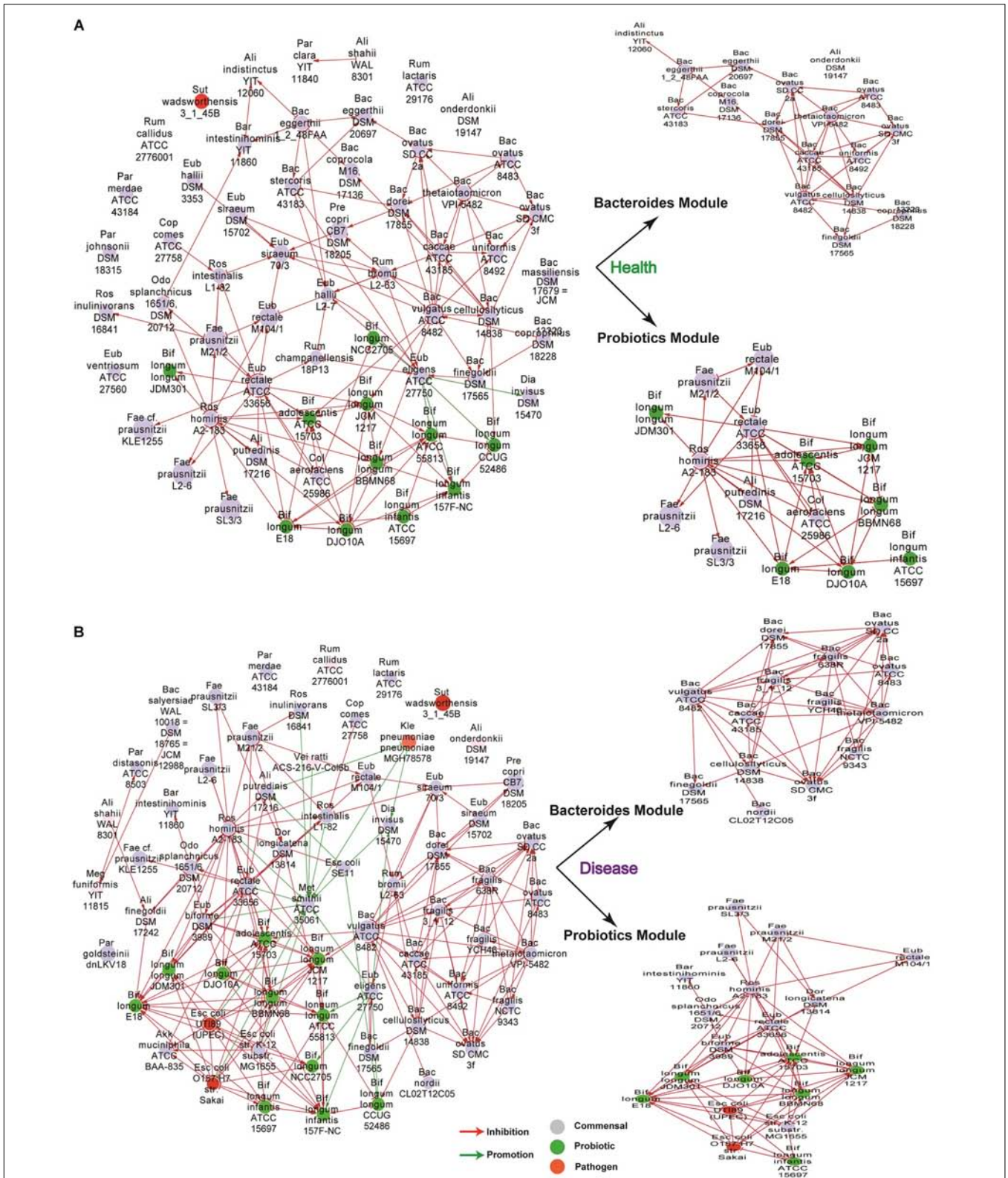


FIGURE 1 | The bacteria interaction networks (weight > 4) obtained from healthy controls **(A)** and patients **(B)**. ForceAtlas2 layout in Gephi (Bastian et al., 2009) was used for this representation. Nodes filled with gray, green, red, and pink represent commensal, probiotic, pathogenic, and opportunistic pathogenic bacteria, respectively. Two main subclusters were identified, one includes mostly probiotic bacteria (Probiotics Module), while the other consists mostly of species in the genus bacteroides (Bacteroides Module).

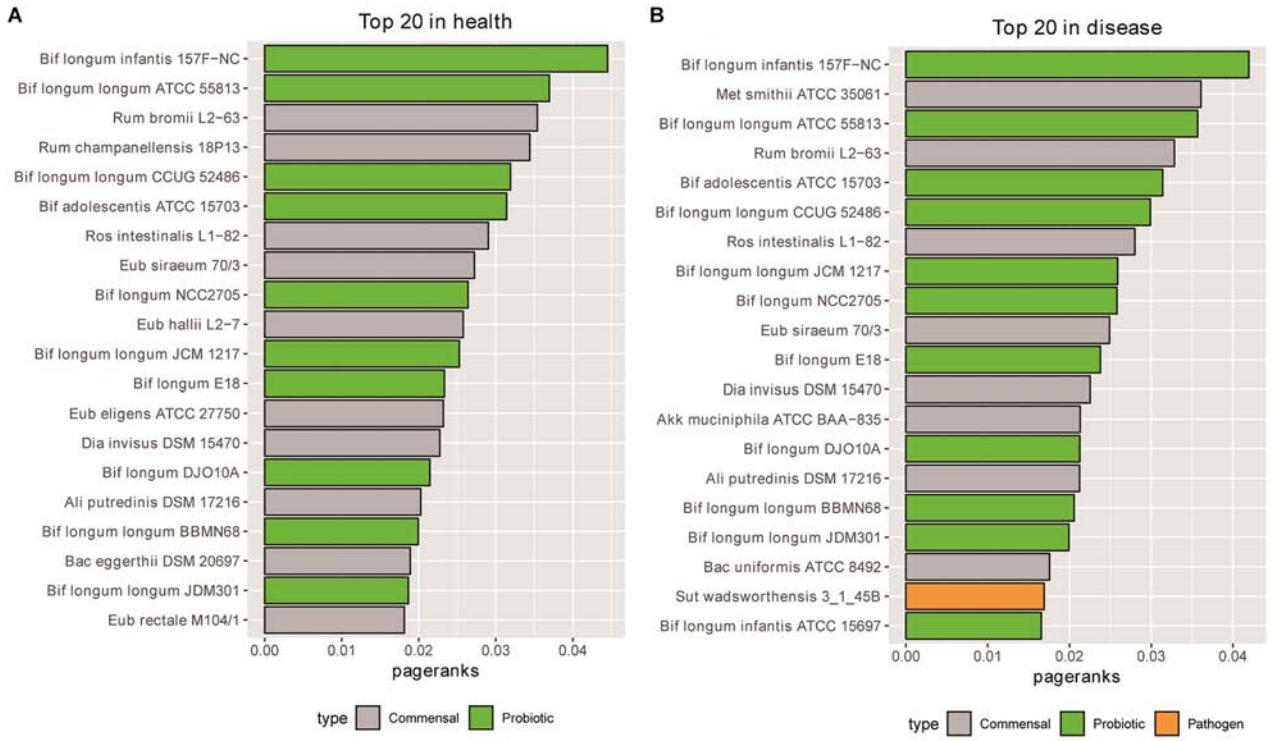


FIGURE 2 | Most of the top 20 bacteria based on PageRanks are probiotics in health **(A)** and disease **(B)** states.

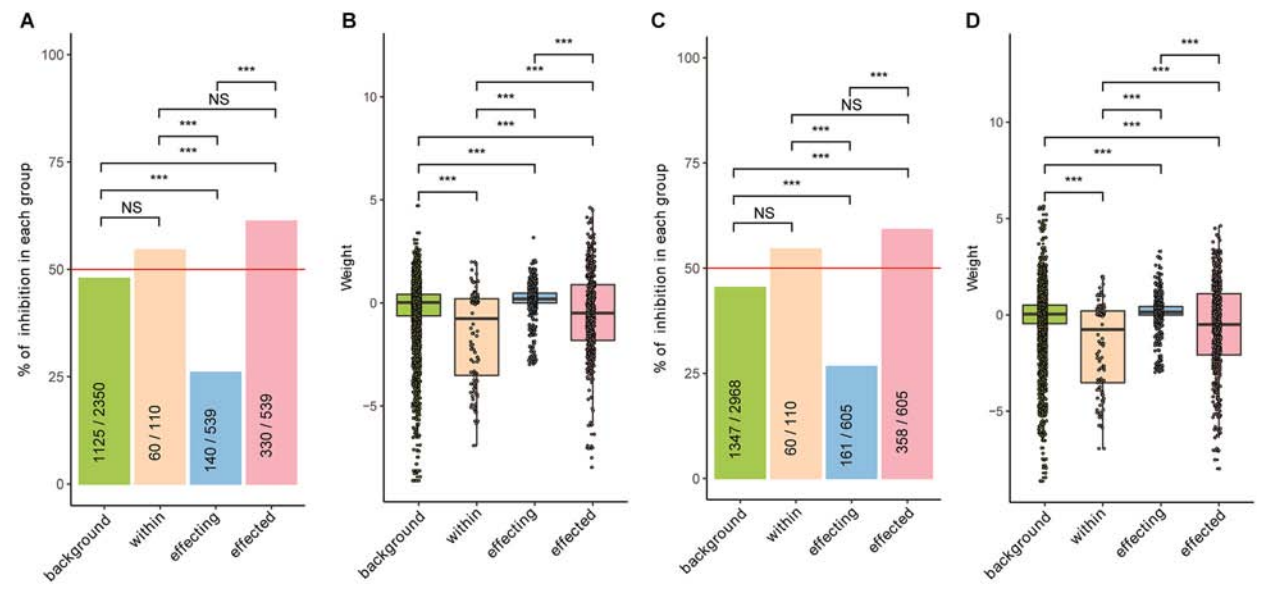


FIGURE 3 | The growth of probiotics was strongly inhibited by other bacteria in both patients and healthy controls. **(A,C)** Proportions of inhibitory interactions in the four groups, calculated separately for patients **(A)** and healthy controls **(C)**; Chi-square test was used to test pairwise differences between two groups. **(B,D)** Distribution of weight values in the four groups, calculated separately for patients **(B)** and healthy controls **(D)**; Wilcoxon Rank Sum test was used for pairwise comparisons between two groups. Interaction data of the four groups are: background – interactions among bacteria excluding probiotics; within – interactions among probiotics; affecting – impacts of probiotics on others; affected – impacts of others on probiotics. Level of significance: NS – not significant; ***$p < 0.01$.

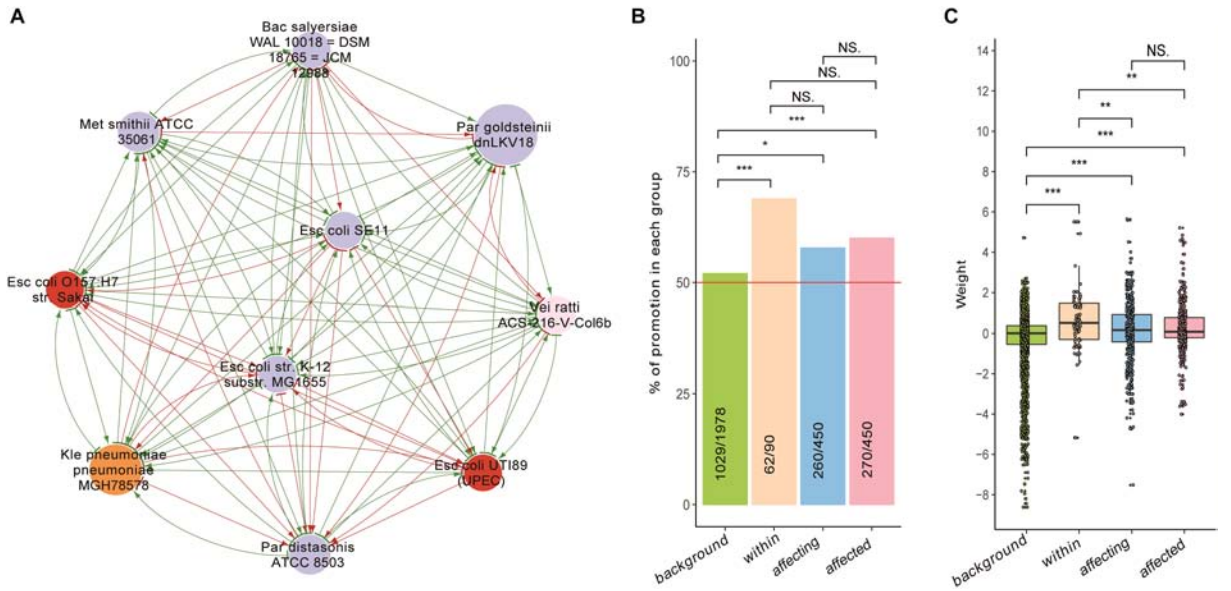


FIGURE 4 | The growth of disease-enriched bacteria could be promoted by themselves and others. **(A)** An exemplary network of disease-enriched bacteria. **(B)** Proportion of promoting interactions in each group; chi-square test was used to perform pairwise (two-groups at a time) comparisons. **(C)** Distribution of weight values in the four groups; Wilcoxon Rank Sum test was used for pairwise comparisons. Interaction data of the four groups are: background – interactions among bacteria excluding probiotics and disease-enriched ones; within – interactions among disease-enriched bacteria; affecting – impacts of disease-enriched bacteria on others; affected – impacts of others on disease-enriched bacteria. Level of significance: NS – not significant; $*p < 0.05$; $**0.01 < p < 0.05$; $***p < 0.01$.

Alterations of the Gut Microbiota During Enteropathogenesis Can Be Explained by Their Immediate Neighbors in the Metabolic Dependency Network

We next checked if alterations of the gut bacteria could be explained by their immediate neighbors in the network. For a given node (species) in the network, we considered two parameters in this calculation, namely the weight of the interactions (w) and the relative abundances (a) of its connecting nodes, and calculated an Inbound Influence Index using following equation: $\sum(\bar{w} \times a)$. As shown in **Table 1** and **Supplementary Table S5**, we were able to predict up to 75% of the directions (i.e., increase or decrease) of the nodes in the metabolic dependency network.

DISCUSSION

In this study, we constructed metabolic dependency networks using gut microbiota datasets of common entero-diseases including IBD and CRC, and revealed unappreciated interaction patterns of disease-enriched bacteria and probiotics. In addition, we showed that the alterations of the gut microbiota during enteropathogenesis can be explained by their immediate neighbors in the metabolic dependency network with reasonable accuracy.

We used Wilcoxon Rank Sum test to identify differentially abundant species between patients and healthy controls. Although the identified significantly changed bacteria are quite different in the two IBD datasets (both contained patients and healthy controls, see **Supplementary Table S1**), we found similar interaction patterns ("mutual inhibition" between probiotics and "mutual promotion" between those significantly enriched bacteria) in the two IBD datasets and the CRC dataset.

Here, the classification of the bacteria (Commensal, Pathogen, and Probiotic) is provided by the literature (Magnusdottir et al., 2017), which is shown in **Supplementary Table S2**. Some strains in *Bifidobacterium bifidum*, which belong to the probiotics, were identified as the most variable strains between the healthy and disease. It is generally known that probiotics can improve human health. A precise definition of probiotics has been proposed by Laurent Verschuere (Verschuere et al., 2000). It was defined as a live microbial adjunct which has a beneficial effect on the host by modifying the host-associated or ambient microbial community, by enhancing the host response toward disease, by improving the quality of its ambient environment, or by ensuring improved use of the feed or enhancing its nutritional value. Above all, the most commonly purported benefits of the consumption of probiotics is modulation of host immunity (Corthésy et al., 2007). Because of these merits, the market for probiotics and probiotic-containing commercial products is constantly growing (Marco et al., 2006; Varankovich et al., 2015). However, a stable microbial community cannot be achieved by a sudden increase in nutrients due to exogenous feeding with probiotics (Verschuere et al., 2000). And we report here for the first time that there is a tendency of mutual restrain between the probiotic bacteria. Therefore, it is very important to take the mutual interaction of probiotics into consideration when develop probiotics or "microbial based therapies."

With the growing recognition of the profound impacts of gut microbiota on human health, it is urgent to understand the molecular basis underlying the alterations of individual species in this complex microbial ecosystem. Compared to the undirected co-occurrence network, the metabolic dependency network is directional and thus could provide mechanistic insights into interspecies interactions. Numerous previous studies have suggested that host genetic and environmental factors can influence the diversity and composition of the gut microbiota (Benson et al., 2010). Among the environmental factors, dietary habits has proven to play a dominant role over other possible variables such as geography, climate, sanitation, hygiene, and ethnicity in shaping the gut microbiota (De Filippo et al., 2010; Walker et al., 2011). Our results indicate that at least in part, the alterations of the gut microbiota under different healthy statuses of the hosts, could be attributed to internal factors including species-species interactions of the gut microbes.

Using metabolic interaction network based on pair-wise metabolic dependencies, we found that unappreciated interaction patterns of between-species metabolic interactions could underlie alterations in gut microbiota during enteropathogenesis, and between probiotics and other microbes. Our methods provided a new framework for studying interactions in gut microbiome and their roles in health and disease. Though carefully evaluated, our results are still highly predictive and to be experimentally validated in the future.

AUTHOR CONTRIBUTIONS

W-HC and DD designed the study. TW and SW collected the data. DD analyzed the data and wrote the first draft of the manuscript. W-HC and NG contributed to the manuscript revisions. All authors approved the submission.

FUNDING

This study was financially supported by the National Natural Science Foundation of China (NSFC) (Grant No. 81803850).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01205/full#supplementary-material

REFERENCES

Aarnoutse, R., de Vos-Geelen, J. M. P. G. M., Penders, J., Boerma, E. G., Warmerdam, F. A. R. M., Goorts, B., et al. (2017). Study protocol on the role of intestinal microbiota in colorectal cancer treatment: a pathway to personalized medicine 2.0. *Int. J. Colorectal. Dis.* 32, 1077–1084. doi: 10.1007/s00384-017-2819-3

Andrews, S. (2014). *FastQC: a Quality Control Tool for High Throughput Sequence Data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed December 1, 2018).

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks in *Proceedings of the Third International ICWSM Conference*, (San Jose: ICWSM).

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18933–18938. doi: 10.1073/pnas.1007028107

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

Cardinale, M., Grube, M., Erlacher, A., Quehenberger, J., and Berg, G. (2015). Bacterial networks and co-occurrence relationships in the lettuce root microbiota. *Environ. Microbiol.* 17, 239–252. doi: 10.1111/1462-2920.12686

Chen, P., Xie, H., Maslov, S., and Redner, S. (2007). Finding scientific gems with google's pagerank algorithm. *J. Informetr.* 1, 8–15. doi: 10.1016/j.joi.2006.06.001

Chow, C. E., Kim, D. Y., Sachdeva, R., Caron, D. A., and Fuhrman, J. A. (2014). Top-down controls on bacterial community structure: microbial network analysis of bacteria. T4-like viruses and protists. *ISME J.* 8, 816–829. doi: 10.1038/ismej.2013.199

Chu, H., Khosravi, A., Kusumawardhani, I. P., Kwon, A. H. K., Vasconcelos, A. C., Cunha, L. D., et al. (2016). Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. *Science* 352, 1116. doi: 10.1126/science.aad9948

Corthésy, B., Gaskins, H. R., and Mercenier, A. (2007). Cross-talk between probiotic bacteria and the host immune system. *J. Nutr.* 137, 781S–790S. doi: 10.1093/jn/137.3.781S

Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 86–92. doi: 10.1038/nature02456

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

Delzenne, N. M., Cani, P. D., Everard, A., Neyrinck, A. M., and Bindels, L. B. (2015). Gut microorganisms as promising targets for the management of type 2 diabetes. *Diabetologia* 58, 2206–2217. doi: 10.1007/s00125-015-3712-7

Drosos, I., Tavridou, A., and Kolios, G. (2015). New aspects on the metabolic role of intestinal microbiota in the development of atherosclerosis. *Metabolism* 64, 476–481. doi: 10.1016/j.metabol.2015.01.007

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

NIH HMP Working Group, N. I. H., Working Group, N. H. W., Peterson, J., Garges, S., Giovanni, M., Mcinnes, P., et al. (2009). The NIH human microbiome project. *Genome Res.* 19, 2317–2323. doi: 10.1101/gr.096651.109

Jackson, M. A., Verdi, S., Maxan, M. E., Shin, C. M., Zierer, J., Bowyer, R. C. E., et al. (2018). Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat. Commun.* 9, 2655. doi: 10.1038/s41467-018-05184-7

Joossens, M., Huys, G., Cnockaert, M., De Preter, V., Verbeke, K., Rutgeerts, P., et al. (2011). Dysbiosis of the faecal microbiota in patients with crohn's disease and their unaffected relatives. *Gut* 60, 631–637. doi: 10.1136/gut.2010.223263

Kim, J., and Hastak, M. (2018). Social network analysis: characteristics of online social networks after a disaster. *Int. J. Inform. Manag.* 38, 86–96. doi: 10.1016/j.ijinfomgt.2017.08.003

Kong, W., Meldgin, D. R., Collins, J. J., and Lu, T. (2018). Designing microbial consortia with defined social interactions. *Nat. Chem. Biol.* 14, 821–829. doi: 10.1038/s41589-018-0091-7

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., et al. (2011). The european nucleotide archive. *Nucleic Acids Res.* 39, D28–D31. doi: 10.1093/nar/gkq967

Liang, Q., Chiu, J., Chen, Y., Huang, Y., Higashimori, A., Fang, J., et al. (2017). Fecal bacteria act as novel biomarkers for noninvasive diagnosis of colorectal cancer. *Clin. Cancer Res.* 23, 2061–2070. doi: 10.1158/1078-0432.CCR-16-1599

Magnusdottir, 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, 81–89. doi: 10.1038/nbt.3703

Marco, M. L., Pavan, S., and Kleerebezem, M. (2006). Towards understanding molecular modes of probiotic action. *Curr. Opin. Biotechnol.* 17, 204–210. doi: 10.1016/j.copbio.2006.02.005

Matsuoka, K., and Kanai, T. (2015). The gut microbiota and inflammatory bowel disease. *Semin. Immunopathol.* 37, 47–55. doi: 10.1007/s00281-014-0454-4

Miyoshi, J., and Chang, E. B. (2017). The gut microbiota and inflammatory bowel diseases. *Trans. Res.* 179, 38–48. doi: 10.1016/j.trsl.2016.06.002

Moreno-Indias, I., Cardona, F., Tinahones, F. J., and Queipo-Ortu no, M. I. (2014). Impact of the gut microbiota on the development of obesity and type 2 diabetes mellitus. *Front. Microbiol.* 5:190. doi: 10.3389/fmicb.2014.00190

Noecker, C., Eng, A., Srinivasan, S., Theriot, C. M., Young, V. B., Jansson, J. K., et al. (2016). Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* 1:e00013–e15. doi: 10.1128/mSystems.00013-15

Russo, E., Bacci, G., Chiellini, C., Fagorzi, C., Niccolai, E., Taddei, A., et al. (2018). Preliminary comparison of oral and intestinal human microbiota in patients with colorectal cancer: a pilot study. *Front. Microbiol.* 8:2699. doi: 10.3389/fmicb.2017.02699

Sartor, R. B., and Wu, G. D. (2017). Roles for intestinal bacteria, viruses, and fungi in pathogenesis of inflammatory bowel diseases and therapeutic approaches. *Gastroenterology* 152, 327.e–339.e. doi: 10.1053/j.gastro.2016.10.012

Selber-Hnatiw, S., Rukundo, B., Ahmadi, M., Akoubi, H., Al-Bizri, H., Aliu, A. F., et al. (2017). Human gut microbiota: toward an ecology of disease. *Front. Microbiol.* 8:1265. doi: 10.3389/fmicb.2017.01265

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

Shrestha, R. K., Lubinsky, B., Bansode, V. B., Moinz, M. B., Mccormack, G. P., and Travers, S. A. (2014). QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics* 15:33. doi: 10.1186/1471-2105-15-33

Sommer, F., Anderson, J. M., Bharti, R., Raes, J., and Rosenstiel, P. (2017). The resilience of the intestinal microbiota influences health and disease. *Nat. Rev. Microbiol.* 15, 630–638. doi: 10.1038/nrmicro.2017.58

Tai, N., Wong, F. S., and Wen, L. (2015). The role of gut microbiota in the development of type 1, type 2 diabetes mellitus and obesity. *Rev. Endocr. Metab. Disord.* 16, 55–65. doi: 10.1007/s11154-015-9309-0

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

Tomasello, G., Sorce, A., Damiani, P., Sinagra, E., and Carini, F. (2017). The importance of intestinal microbial flora (microbiota) and role of diet. *Prog. Nutr.* 19, 342–344. doi: 10.23751/pn.v19i3.4495

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, 902–903. doi: 10.1038/nmeth.3589

van der Ark, K. C. H., van Heck, R. G. A., Martins Dos Santos, V. A. P., Belzer, C., and de Vos, W. M. (2017). More than just a gut feeling: constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes. *Microbiome* 5, 78. doi: 10.1186/s40168-017-0299-x

Varankovich, N. V., Nickerson, M. T., and Korber, D. R. (2015). Probiotic-based strategies for therapeutic and prophylactic use against multiple gastrointestinal diseases. *Front. Microbiol.* 6:685. doi: 10.3389/fmicb.2015.00685

Verschuere, L., Rombaut, G., Sorgeloos, P., and Verstraete, W. (2000). Probiotic bacteria as biological control agents in aquaculture. *Microbiol. Mol. Biol. Rev.* 64, 655–671. doi: 10.1128/mmbr.64.4.655-671.2000

Walker, A. W., Ince, J., Duncan, S. H., Webster, L. M., Holtrop, G., Ze, X., et al. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* 5, 220–230. doi: 10.1038/ismej.2010.118

Yamashita, T., Kasahara, K., Emoto, T., Matsumoto, T., Mizoguchi, T., Kitano, N., et al. (2015). Intestinal immunity and gut microbiota as therapeutic targets for preventing atherosclerotic cardiovascular diseases. *Circ. J.* 79, 1882–1890. doi: 10.1253/circj.CJ-15-0526

Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D. R., Bork, P., and Patil, K. R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl. Acad. Sci. U.S.A.* 112, 6449–6454. doi: 10.1073/pnas.1421834112

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

Zhang, C., and Hua, Q. (2016). Applications of genome-scale metabolic models in biotechnology and systems medicine. *Front. Physiol.* 6:413. doi: 10.3389/fphys.2015.00413

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 Statement: 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.

Copyright © 2019 Dai, Wang, Wu, Gao and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 03 July 2019
doi: 10.3389/fmicb.2019.01467

Assessing the Hybrid Effects of Neutral and Niche Processes on Gut Microbiome Influenced by HIV Infection

Guanshu Yin[1] *and Yao Xia*[2]*

[1] *The Second Affiliated Hospital of Kunming Medical University, Kunming, China,* [2] *Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China*

That both stochastic neutral and deterministic niche forces are in effect in shaping the community assembly and diversity maintenance is becoming an increasingly important consensus. However, assessing the effects of disease on the balance between the two forces in the human microbiome has not been explored to the best of our knowledge. In this article, we applied a hybrid model to address this issue by analyzing the potential effect of HIV infection on the human gut microbiome and adopted a further step of multimodality testing to improve the interpretation of their model. Our study revealed that although niche process is the dominant force in shaping human gut microbial communities, niche process- and neutral process-driven taxa could coexist in the same microbiome, confirming the notion of their joint responsibility. However, we failed to detect the effect of HIV infection in changing the balance. This suggests that the rule governing community assembly and diversity maintenance may be changed by the disturbance from HIV infection-caused dysbiosis. Although we admit that the general question of disease effect on community assembly and diversity maintenance may still be an open question, our study presents the first piece of evidence to reject the significant influence of diseases.

Keywords: neutral theory, niche theory, microbiome analyses, hybrid model, HIV

OPEN ACCESS

Edited by:
Matthias Hess,
University of California, Davis,
United States

Reviewed by:
Joseph P. De Santis,
University of Miami, United States
Grégory Dubourg,
IHU Méditerranée Infection, France

***Correspondence:**
Yao Xia
xiayao0125@outlook.com

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

Received: *25 March 2019*
Accepted: *11 June 2019*
Published: *03 July 2019*

Citation:
Yin G and Xia Y (2019) Assessing the Hybrid Effects of Neutral and Niche Processes on Gut Microbiome Influenced by HIV Infection. Front. Microbiol. 10:1467. doi: 10.3389/fmicb.2019.01467

INTRODUCTION

Human gut is an ideal micro ecosystem colonized by countless microbes, where the mechanistic explanation of species abundance distribution (SAD) needs to be clarified. Typically, the forces that shape and maintain the biodiversity of community are thought to be controlled by deterministic factors, such as host species, genotype, diet, health, competition and niche differentiation, which has been referred to *Niche Theory*, but it fails to explain a number of rare taxa could coexist in very diverse environments when applied to macro-organisms (Ofiteru et al., 2010; Burns et al., 2015). The *Neutral Theory* proposed by Hubbell and widely used in the macro-ecology area nowadays has challenged this view. This theory considers trophically similar species are functionally equivalent and the SAD patterns in community can be explained by stochastic processes (Hubbell, 2001, 2006). The *Neutral Theory* combines neutrality, stochasticity, sampling and dispersal and presents a simple null model to test the mechanism of community assembling and biodiversity maintenance

in ecological communities, but its abundance-based simplicity also arises query that is was thought by someone not robust enough, because many parameters may cause similar result (Alonso et al., 2006; Ofiteru et al., 2010; Li and Ma, 2016). Although contentious debates on both theories have been provoked, there have been adequate evidences supporting the idea that neither of them alone is sufficiently to explain the full range of observed SAD in natural communities (Stokes and Archer, 2010). The hypothesis that two contrasting theories are probably jointly responsible for the community assembly (Gravel et al., 2006; Leibold and McPeek, 2006; Zhang et al., 2009; Dumbrell et al., 2010; Stegen et al., 2012) seems more reasonable than either theory alone.

The ecological theories accounting for the community assembly have been tested in macro community widely until now, and their uses on microbial community have been gradually recognized recently, although there is still a long way to completely understand the microbial community. Schmidt et al. (2015) revealed that deterministic processes drive fish microbiome assembly dominantly while no evidence supports stochastic theory. O'Dwyer et al. (2015) also tested neutral theory for several datasets to determine whether or not the observed SAD patterns can fit the neutral prediction and found a clear departure from the predictions of standard neutral theories, indicating that standard neutral models may not provide the most useful null models for microbial communities. Li and Ma (2016) tested more than 7000 samples from different parts of human body with neutral theory, which revealed that very few microbial communities passed the neutral prediction. Given the hypothesis of hybrid effects of neutral and niche theory, it is too arbitrary to draw a conclusion that neutral processes do not play a role in microbial community. In this study, with the aim to identify the possible neutral process within non-neutral microbial communities, we used a hybrid model considering both neutral and niche theory to test a dataset of human gut sample. As the dataset we used contains samples from healthy individuals and patients with HIV that is proven to cause dramatic dysbiosis of gut microbiome (Lozupone et al., 2013; Mutlu et al., 2014; Dubourg et al., 2016; Williams et al., 2016), our another aim is to recognize the alteration of the mechanism of gut microbiome assembly resulted from HIV infection.

As mentioned before, it is difficult to identify quantitatively the exact processes shaping community from certain SAD patterns with standard neutral theories empirically. Quantifying the relative roles of neutral and niche process brings a non-trivial task. Microbial community is usually characterized by rich biodiversity, suggesting a high possibility that at least subpopulations controlled mainly by neutral process exist in a community. In a global view, such community consisting of both neutrally assembly and niche-selected taxa may not be recognized as a neutral community *via* neutral models concerning SAD pattern alone. Hence it is necessary to adopt more information, such as phylogenetic analysis, to build sophisticated models to evaluate the relative roles of both processes. Jeraldo et al. (2012) proposed a model that fuses measures of abundance with phylogenetic information that has been attracting increasingly attention in ecological studies (Kelly et al., 2008; Cavender-Bares et al., 2009; Jabot and Chave, 2009; Burbrink et al., 2015) to address this problem, which is particularly suitable for microbial community characterized by high level of biodiversity. The authors identified successfully subpopulations of the chicken gastrointestinal tracts that may be undergoing neutral process *via* the observation of a small non-zero peak within the distance-based plot generated by their genomic-based model. However, observations from histograms may sometimes be hard to determine and even misrepresent the real distributions. Therefore, we went a further step by adopting a multimodality statistical test, Silverman (1981), to improve the interpretation of Jeraldo et al. (2012) model.

MATERIALS AND METHODS

Dataset Reprocessing

In Lozupone et al. (2013) study, they collected fecal samples of individuals with chronic HIV infection ($n = 22$), recent infection ($n = 3$), and HIV-negative controls ($n = 13$) on sterile swabs either during or 12 hr prior to the visit and then all individuals were assigned into four cohorts: (1) recent HIV-1 infection, individuals likely infected within the prior 6 months; (2) chronic HIV-1 infection untreated, individuals infected for >6 months and ART drug naive or off treatment for >6 months; (3) chronic HIV-1 infection on long-term ART, ART treatment for ≥12 months with a minimum of three ART drugs prior to study entry and viral suppressed for >6 months; and (4) healthy controls. Some individuals donated samples twice and 58 samples entered into the 16s rRNA sequencing process finally. The reads for each sample were stored at EBI[1] (Accession Number ERP003611). In our study, we downloaded the sequencing data from EBI and selected qualified samples with enough numbers of high quality sequences to perform the recalculation.

We used Jeraldo et al. (2012) pipeline (Tornado[2]) to reanalyze the reads data. In brief, short sequences (shorter than 100 bp) and chimeras were eliminated firstly and remaining sequences were aligned using Mothur and Silva reference database. To make sure the sequences start and end at the same position, the ends of all alignments were trimmed. Then operational taxonomic units (OTUs) were picked with the complete linkage method of Schloss et al. (2009) with a cutoff of 3% sequence identity. All reprocessed OTUs entered the following analysis.

The Computational Procedure for the Hybrid Model

For each sample, the observed OTUs can be classified into two categories: modal OTUs (most abundant) and rare OTUs (less abundant), on the basis of a threshold value k. The core idea is to visualize the correlations between modal and rare OTUs, which will depend on the ecological dynamics, using the information obtained from the phylogenetic distances of representative sequences between both types of OTUs and the abundances of OTUs in a hypothetic high-dimensional sequence space.

[1]http://www.ebi.ac.uk/ena/

[2]http://tornado.igb.uiuc.edu/bio/

In the first case, suppose that a community is drove by neutral process, hence modal and rare OTUs in this community would distribute at random in the hypothetic space. The distances between OTUs (representative sequences) can be measured using a normalized Hamming distance:

$$H_{ij} = \frac{1}{L}\sum_{\alpha=1}^{L}[1 - \delta(S_\alpha^i - S_\alpha^j)] \quad (1)$$

where H is the distance between the ith and jth OTU, L is the length of the representative sequence (to compare the Hamming distance, all the analyzed sequences should be in the same length), δ is the Kronecker delta and S represents the label of base at a given position (from α to L) in the sequence with superscripts (e.g., i and j) indicating OTUs. S takes values 1, 2, 3, 4 corresponding to the four bases *ACGT*.

Ideally, in a neutral process-driven community, the mean of H should be 3/4 as the chance that two bases at the same position are identical is 1/4. However, considering the complications deriving from highly conserved bases that cannot be appropriately modeled as being chosen randomly from the alphabet in reality, the actual value of the mean of H would be *3(L-M)/4L*, given there are M conserved bases in the sequence. Then, for each rare OTU (labeled k), the distances between it and all modal OTUs are calculated *via* method described above and the shortest one is selected and labeled E_k. $\{E_k\}$ is a subset of $\{H_{ij}\}$ and its distribution is also a bell-shape plot peaked at a slightly smaller value than mean $\{H_{ij}\}$.

In another case, i.e., when the community is driven by niche process, the distributions of both types of OTUs are not random, where rare OTUs that evolve from modal OTUs through a few point mutations surround the modal OTUs closely. Such distributions are also observed in Jeraldo et al. (2012) study using a weighted version of principal component analysis (PCA) to reduce the hypothetic space into a 2D space, which are obviously different from the distributions in neutral community. All the normalized H_{ij} from each rare OTU to the nearest modal OTU are calculated *via* making a Voronoi polyhedron construction in the hypothetic space. Thus, the probability distribution of E_k should be a delta distribution that is peaked at $E = 0$ and decreases monotonically for $E > 0$.

In most cases, both neutral and niche process will not be responsible for the construction of the community solely therefore the hybrid effect should be took into account. Jeraldo et al. (2012) evaluated the hybrid effect of model when there are αN OTUs undergo a niche dynamic in a community containing N OTUs using Monte Carlo simulations on a simulated dataset. The parameter α can serve as an indicator suggesting observed community is driven by purely neutral process ($\alpha = 0$), purely niche process ($\alpha = 1$) or hybrid process (α is between 0 and 1). The distributions of distance plot in three types of community are clearly distinct in shape: for purely neutral community, the distribution plot is bell-shaped; for purely niche community, the distribution is a delta distribution that is peaked at $E = 0$ and then decreases for $E > 0$ (the authors found for niche-like models, the peak at zero moves to a non-zero peak that corresponds to the average size of the niche); for hybrid community, the distribution plot shows characteristic of niche distribution at the first-member (i.e., starts at a non-zero peak and then decreases monotonically) and characteristic of neutral distribution at the end-member (i.e., a non-zero peak arises in the end-member). The non-zero peak at the end-member can be viewed as an evidence for the presence of subpopulations shaped mainly by neutral process and the sequences within the peak may be undergoing neutral dynamics.

Jeraldo et al. (2012) also tried to improve their model by weighting the contribution of E_k by abundance of OTU but failed to find change in the neutral community and qualitative change in the niche community, suggesting the distribution of distance may be only weakly dependent on the abundance distribution of OTUs. To simplify the calculation, the information of OTU abundance is not included in such analysis. Another concern is the choice of the threshold k. In the original study, the authors revealed that the results of the metric on model systems are unchanged when k is changing between 2 and 10% hence they select 1 to 10% for this parameter in their pipeline. In our study, we set the k value as 5%, which is used by the authors to present their experimental data in the original article. The complete computational procedure was performed in the softer ware Tornado see text footnote 2.

Multimodality Test for the Distribution of Nearest Distances

The mode of a distribution is the value having highest probability of being observed. On the basis of the principle of their model, the non-zero peak at the right side of the histogram of nearest distance represents the neutral-driven subpopulations in a community. Thus, a pure niche process-driven community should have only one mode, and community driven by niche-neutral hybrid process should have more than one mode. Given the pure neutral community did not exist in the real world, the unimodality should represents niche process and multimodality should be equivalent to hybrid process. Silverman (1981) provides a classical method to test the null hypothesis that a distribution has at most n modes ($n = 1$ in our study), where the alternative hypothesis that the distribution has more than n modes can be rejected on the basis of a p-Value. In this study, for each sample, we draw both histogram and kernel density plot using the distribution of distance to the nearest modal OTU of every rare OTU in the sample community with R (version 3.3.2) and conducted Silverman (1981) with the R package, *silvermantest*[3].

RESULTS AND DISCUSSION

Only samples with enough numbers of high quality sequences were included in our study. In total, 52 qualified samples with 4 in cohort (1), 18 in cohort (2), 8 in cohort (3), and 21 in cohort (4) were selected. As the number of individuals in four cohorts are not balance, we here assigned cohort 1, 2, and 3 into HIV group and cohort 4 into Non-HIV group

[3]https://www.mathematik.uni-marburg.de/~stochastik/R_packages/

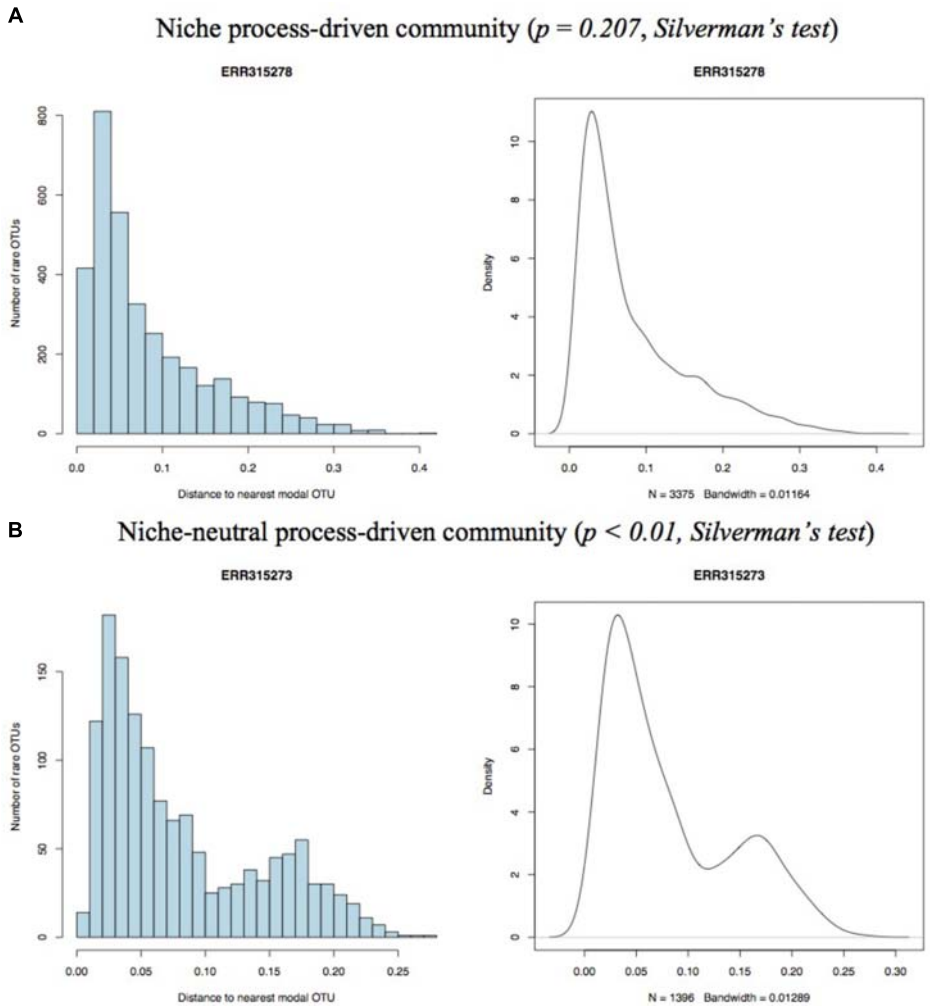


FIGURE 1 | Representative histograms (left) and kernel density plots (right) for niche process- **(A)** and niche-neutral process-driven **(B)** community.

and compared the number of samples that fitted neutral-niche plot between two groups. For each sample, we generated both histogram and density plot on the distance of each rare OTU to nearest modal OTU and performed Silverman's test for multimodality. Two pairs of representative histograms and density plots from niche process- and hybrid process-driven community, respectively, are displayed in **Figure 1**. The remains are shown in **Supplementary Figure S1** and results of Silverman's test are listed in **Supplementary Table S1**. The number of samples passing Silverman's test for each cohort and group are displayed in **Table 1**. In detail, 56.86% (29/51) samples in total passed the Silverman's test ($p < 0.05$), indicating they are undergoing niche-neutral process, thus 43.14% (22/51) samples are driven by niche process dominantly. According to different cohorts, 25% (1/4) samples in cohort (1), 72.22% (13/18) samples in cohort (2), 75% (6/8) samples in cohort (3) and 42.68% (9/21) samples in cohort (4) passed the multimodality test. No one sample satisfied the pure neutral plot, so the remains in each cohort are niche-driven community, suggesting that although niche process is the dominant force in shaping gut microbiome assembly and play roles in all samples, the contributions made by neutral process should not be neglected.

TABLE 1 | The number of samples passing Silverman's test.

Group	Cohort	Samples driven by niche-neutral process	Samples driven by niche-neutral process
HIV infection	Cohort 1	25% (1/4)	66.67% (20/30)
	Cohort 2	72.22% (13/18)	
	Cohort 3	75% (6/8)	
Non-HIV infection	Cohort 4	42.86% (9/21)	42.86% (9/21)

Because the number of individuals in four cohorts are not balance, we than assigned cohort 1, 2, and 3 into HIV group and cohort 4 into Non-HIV group and compared the number of samples that fitted neutral-niche plot between two groups. In HIV group, 66.67% (20/30) samples fitted neutral-niche

plots and 42.86% (9/21) fitted neutral-niche plots in Non-HIV group. There is no significant difference between the results of two groups (p = 0.1504, *Fisher's exact test*), indicating HIV infection may not change the force that shapes gut microbiome assembly essentially.

Four representative neutral-niche plots and niche-like plots for each cohort are displayed in **Figure 1** and remaining plots can be found in **Supplementary Figure S1**.

It has been generally accepted that the community assembly is an important topic in the macro-community ecology area and can be accounted for by different ecological theories (Niu et al., 2009; Latombe et al., 2015). From the view of ecologists, however, the microbial community where the birth, death, immigration and speciation happen at any time is also an important subject that is controlled by general ecological principles and laws. Hence the two important theories explaining the assembly of community, niche theory underlying deterministic factors and neutral theory underlying stochastic factors, should be also appropriate to describe the microbial community assembly.

In this study, we performed a test of hybrid model considering both neutral and niche process on human gut samples and did not find one sample satisfy the prediction of pure neutral theory, which is consistent with our previous report, where we tested Human Microbiome Project (HMP) dataset with neutral models and found none of the gut samples passed the neutrality test (Li and Ma, 2016). The reason may be the high level of diversity and complexity of organisms colonizing in gut, just as the study by Fisher and Mehta (2014) suggests that the communities with large population sizes and relatively stable environment is more like driven mainly by niche-process and the communities with small population size and unstable environment is more likely driven by neutral-process. Fisher and Mehta (2014) also reveals the presence of a phase transition process between niche-driven phases and neutral-driven phases in communities, suggesting a neutral-niche phase where niche process and neutral process are jointly responsible for the assembly and maintenance of community should exist in amount of communities. Although our study confirmed that niche process play its role dominantly in shaping gut microbiome, our results also provides evidence that there are subpopulations driven mainly by neutral process in certain overall neutral communities, indicated by the observation of neutral-niche plots containing a small non-zero peak comparing with niche-like plots, where the sequences under the non-zero peak should be belong to the neutral-driven taxa. The coexist of neutral and niche process has been supported by many studies (Leibold and McPeek, 2006; Dumbrell et al., 2010; Ofiteru et al., 2010; Ayarza and Erijman, 2011), which may result from different physical reasons (Jeraldo et al., 2012), such as the fact that those generalist microbes that can exist in various environments constitute the neutrally assembly part of microbiome (Langenheder and Szeìkely, 2011), while the niche portion consists of those specific microbes that is adapted to the medium conditions (Burke et al., 2011). Disturbances of gut physiological environment may cause selections on microbes, especially for those niche process-driven taxa, hence we also wonder the question that whether or not the typical dysbiosis caused by HIV infection is linked to change of the force that shapes gut microbiome assembly. We compared the gut microbiome assembly forces between HIV infection and health and found that despite the different progress of HIV patients, the proportion of samples fitted niche-neutral plots in HIV group is 66.67% (20/30), which has no significant difference with the proportion of 42.86% (9/21) in health group, implying HIV infection would not change the rule that shapes gut microbiome assembly essentially.

Jeraldo et al. (2012) hybrid model provides a very effective tool to quantify the relative roles of niche and neutral processes shaping microbial community by merging measures of abundance with phylogenetic information, but there is room for improvement. First, it adopts distance-based histogram to measure the neutral-niche process, where the shape of plots sometimes is difficult to distinguish *via* observation, especially when the non-zero peak representing neutral process is not very evident. Second, the smoothness of distance plot is dependent largely on the number of OTUs, hence when the number of OTUs is not large enough, the shape of distance plot would be shapeless so that little information can be achieved. For the first issue, we adopted a statistical method, Silverman's test, for testing multimodality *via* kernel density estimation, through which the niche-neutral hybrid process-driven community can be distinguished from niche community. As to the second issue, according to our experiences, for human microbiome samples, the gut samples usually can satisfy the requirement of the number of OTUs whereas other body parts such as lung and oral can hardly meet the number that generates eligible plots.

In summary, we firstly adopted a Silverman's test on the original results of Jeraldo et al. (2012) hybrid model, which improved the interpretation of the results. Then we use this strategy to reanalyze a dataset of HIV-related human gut microbiome in order to find HIV-specific changes in the assembly of gut microbial community. Our results revealed that although niche process is dominant in shaping human gut microbiome, niche process- and neutral process-driven taxa could coexist in the same microbiome, confirming the idea that niche and neutral processes may be jointly responsible for the gut microbial community assembly and HIV infection-caused dysbiosis may not change the force that shape the assembly of gut microbiome. Besides the evidences that niche and neutral process may co-occurrence in gut microbiome, our study also offer suggestions for improvement of Jeraldo et al. (2012) model *via* introducing statistical multimodality test method.

Our study is a pilot study reanalyzing only one dataset with limited samples. This is the major limitation of our study. As there has not been a consistent HIV-specific dysbiosis pattern of gut microbiome, which reflects from the contradictory results of studies using different samples and datasets (Noguera-Julian et al., 2016). This may because gut microbiome is related to plenty of factors in addition to HIV infection, such as socioeconomic factors, geography, age, diet, drug use, genetic, lifestyle and sex preference (Noguera-Julian et al., 2016; Liu et al., 2017; Williams, 2019). For example, Noguera-Julian et al. (2016) found that the gut microbiome of men who has sex with men (MSM) was richer and more diverse than that of non-MSM men. Given the limited samples available for each study, it is challenging to control all

the confounders in HIV-related gut microbiome study. Thus few studies have successfully established causal links between changes of the gut microbial community composition and HIV infection. Likewise, to investigate the HIV-specific change of assembly of gut microbial community is also a non-trivial task. In further study, we would try to collect more samples from different datasets and use a hierarchical analysis strategy to achieve more reliable results.

DATA AVAILABILITY

The dataset analyzed for this study can be found in the EBI (http://www.ebi.ac.uk/ena/; Accession Number ERP003611).

AUTHOR CONTRIBUTIONS

YX conceived and designed the experiments, prepared the figures and tables, and drafted the manuscript. YX and GY carried out the experiments, analyzed the data, interpreted the results, and approved the final draft of the manuscript. GY reviewed the draft of the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Zhanshan Ma at Kunming Institute of Zoology for providing the valuable advices and revising the manuscript. We also appreciate the computational and testing helps from Ms. Dandan Zeng and of the Computational Biology and Medical Ecology Lab.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01467/full#supplementary-material

REFERENCES

Alonso, D., Etienne, R. S., and Mckane, A. J. (2006). The merits of neutral theory. *Trends Ecol. Evol.* 21, 451–457.

Ayarza, J. M., and Erijman, L. (2011). Balance of neutral and deterministic components in the dynamics of activated sludge floc assembly. *Microb. Ecol.* 61, 486–495. doi: 10.1007/s00248-010-9762-y

Burbrink, F. T., Mckelvy, A. D., Pyron, R. A., and Myers, E. A. (2015). Predicting community structure in snakes on eastern nearctic islands using ecological neutral theory and phylogenetic methods. *Proc. Biol. Sci.* 282:20151700. doi: 10.1098/rspb.2015.1700

Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S., and Thomas, T. (2011). Bacterial community assembly based on functional genes rather than species. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14288–14293.

Burns, A. R., Stephens, W. Z., Stagaman, K., Wong, S., Rawls, J. F., Guillemin, K., et al. (2015). Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME J.* 10, 655–664. doi: 10.1038/ismej.2015.142

Cavender-Bares, J., Kozak, K., Fine, P., and Kembel, S. (2009). The merging of community ecology and phylogenetic biology. *Ecol. Lett.* 12, 693–715. doi: 10.1111/j.1461-0248.2009.01314.x

Dubourg, G., Surenaud, M., Lévy, Y., Hüe, S., and Raoult, D. (2016). Microbiome of HIV-infected people. *Microb. Pathog.* 106, 85–93. doi: 10.1016/j.micpath.2016.05.015

Dumbrell, A. J., Nelson, M., Helgason, T., Dytham, C., and Fitter, A. H. (2010). Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME J.* 4, 337–345. doi: 10.1038/ismej.2009.122

Fisher, C. K., and Mehta, P. (2014). The transition between the niche and neutral regimes in ecology. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13111–13116. doi: 10.1073/pnas.1405637111

Gravel, D., Canham, C., Beaudet, M., and Messier, C. (2006). Reconciling niche and neutrality: the continuum hypothesis. *Ecol. Lett.* 9, 399–409.

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. New Jersey, NJ: Princeton University Press.

Hubbell, S. P. (2006). Neutral theory and the evolution of ecological equivalence. *Ecology* 87, 1387–1398.

Jabot, F., and Chave, J. (2009). Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecol. Lett.* 12, 239–248. doi: 10.1111/j.1461-0248.2008.01280.x

Jeraldo, P., Sipos, M., Chia, N., Brulc, J. M., Dhillon, A. S., Konkel, M. E., et al. (2012). Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9692–9698. doi: 10.1073/pnas.1206721109

Kelly, C., Bowler, M., Pybus, O., and Harvey, P. (2008). Phylogeny, niches, and relative abundance in natural communities. *Ecology* 89, 962–970.

Langenheder, S., and Szeìkely, A. J. (2011). Species sorting and neutral processes are both important during the initial assembly of bacterial communities. *ISME J.* 5, 1086–1094. doi: 10.1038/ismej.2010.207

Latombe, G., Hui, C., and Mcgeoch, M. A. (2015). Beyond the continuum: a multi-dimensional phase space for neutral-niche community assembly. *Proc. Biol. Sci.* 282:20152417. doi: 10.1098/rspb.2015.2417

Leibold, M., and McPeek, M. (2006). Coexistence of the niche and neutral perspectives in community ecology. *Ecology* 87, 1399–1410.

Li, L., and Ma, Z. (2016). Testing the neutral theory of biodiversity with human microbiome datasets. *Sci. Rep.* 6:31448. doi: 10.1038/sre31448

Liu, J., Williams, B., Frank, D., Dillon, S. M., Wilson, C. C., and Landay, A. L. (2017). Inside out: HIV, the gut microbiome, and the mucosal immune system. *J. Immunol.* 198, 605–614. doi: 10.4049/jimmunol.1601355

Lozupone, C. A., Li, M., Campbell, T. B., Flores, S. C., Linderman, D., Gebert, M. J., et al. (2013). Alterations in the gut microbiome associated with HIV-1 infection. *Cell Host Microbe* 14, 329–339. doi: 10.1097/QAD.0000000000000869

Mutlu, E. A., Keshavarzian, A., Losurdo, J., Swanson, G., Siewe, B., Forsyth, C., et al. (2014). A compositional look at the human gastrointestinal microbiome and immune activation parameters in HIV infected subjects. *PLoS Pathog.* 10:e1003829. doi: 10.1371/journal.ppat.1003829

Niu, K. C., Liu, Y. N., Shen, Z. H., He, F. L., and Fang, J. Y. (2009). Community assembly: the relative importance of neutral theory and niche theory. *Biodiv. Sci.* 17, 579–593.

Noguera-Julian, M., Rocafort, M., Guillén, Y., Rivera, J., Casadellà, M., Nowak, P., et al. (2016). Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine* 5, 135–146. doi: 10.1016/j.ebiom.2016.01.032

O'Dwyer, J. P., Kembel, S. W., and Sharpton, T. J. (2015). Backbones of evolutionary history test biodiversity theory for microbes. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8356–8361. doi: 10.1073/pnas.1419341112

Ofiteru, I. D., Lunn, M., Curtis, T. P., Wells, G. F., Criddle, C. S., Francis, C. A., et al. (2010). Combined niche and neutral effects in a microbial wastewater

treatment community. *Proc. Natl. Acad. Sci. U.S.A.* 107, 457–462. doi: 10.1073/pnas.1000604107

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

Schmidt, V. T., Smith, K. F., Melvin, D. W., and Amaralzettler, L. A. (2015). Community assembly of a euryhaline fish microbiome during salinity acclimation. *Mol. Ecol. Ecol.* 24, 2537–2550. doi: 10.1111/mec.13177

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. R. Stat. Soc. Series B* 43, 97–99.

Stegen, J. C., Lin, X., Konopka, A. E., and Fredrickson, J. K. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J.* 6, 1653–1664. doi: 10.1038/ismej.2012.22

Stokes, C. J., and Archer, S. R. (2010). Niche differentiation and neutral theory: an integrated perspective on shrub assemblages in a parkland savanna. *Ecology* 91, 1152–1162.

Williams, B. (2019). Gut microbiome in HIV Infection: overcoming barriers? *Dig. Dis. Sci.* doi: 10.1007/s10620-019-05500-1 [Epub ahead of print].

Williams, B., Landay, A., and Presti, R. M. (2016). Microreview microbiome alterations in HIV infection a review. *Cell Microbiol.* 18, 645–651.

Zhang, Q., Buckling, A., and Godfray, H. (2009). Quantifying the relative importance of niches and neutrality for coexistence in a model microbial system. *Funct. Ecol.* 23, 1139–1147.

Conflict of Interest Statement: 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.

Copyright © 2019 Yin and Xia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 July 2019
doi: 10.3389/fmicb.2019.01578

RWHMDA: Random Walk on Hypergraph for Microbe-Disease Association Prediction

Ya-Wei Niu[1], Cun-Quan Qu[1,2], Guang-Hui Wang[1,2] and Gui-Ying Yan[3]*

[1] School of Mathematics, Shandong University, Jinan, China, [2] Data Science Institute, Shandong University, Jinan, China, [3] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Based on advancements in deep sequencing technology and microbiology, increasing evidence indicates that microbes inhabiting humans modulate various host physiological phenomena, thus participating in various disease pathogeneses. Owing to increasing availability of biological data, further studies on the establishment of efficient computational models for predicting potential associations are required. In particular, computational approaches can also reduce the discovery cycle of novel microbe-disease associations and further facilitate disease treatment, drug design, and other scientific activities. This study aimed to develop a model based on the random walk on hypergraph for microbe-disease association prediction (RWHMDA). As a class of higher-order data representation, hypergraph could effectively recover information loss occurring in the normal graph methodology, thus exclusively illustrating multiple pair-wise associations. Integrating known microbe-disease associations in the Human Microbe-Disease Association Database (HMDAD) and the Gaussian interaction profile kernel similarity for microbes, random walk was then implemented for the constructed hypergraph. Consequently, RWHMDA performed optimally in predicting the underlying disease-associated microbes. More specifically, our model displayed AUC values of 0.8898 and 0.8524 in global and local leave-one-out cross-validation (LOOCV), respectively. Furthermore, three human diseases (asthma, Crohn's disease, and type 2 diabetes) were studied to further illustrate prediction performance. Moreover, 8, 10, and 8 of the 10 highest ranked microbes were confirmed through recent experimental or clinical studies. In conclusion, RWHMDA is expected to display promising potential to predict disease-microbe associations for follow-up experimental studies and facilitate the prevention, diagnosis, treatment, and prognosis of complex human diseases.

Keywords: hypergraph, random walk, microbe, human diseases, association prediction

OPEN ACCESS

Edited by:
George Tsiamis,
University of Patras, Greece

Reviewed by:
Anastasios Chanalaris,
University of Oxford, United Kingdom
Edwin Wang,
University of Calgary, Canada

****Correspondence:***
Guang-Hui Wang
ghwang@sdu.edu.cn

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

Received: *16 January 2019*
Accepted: *25 June 2019*
Published: *10 July 2019*

Citation:
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

INTRODUCTION

Microbes exist in almost all habitats of flora and fauna, including humans. Deeper microbiological insights have indicated more compact associations between humans and their microflora (Sommer and Backhed, 2013). Some microbes are harmless and vital for host health in various manners, such as enhancement of host immunity, improvement of host metabolic capability, and protection of the host against pathogens (Eckburg et al., 2003; Ventura et al., 2009). Over the past few decades,

numerous studies have focused on microbes inhabiting humans (Peterson et al., 2009). For instance, the gut flora are a complicated microbial community in the human digestive tract (Sommer and Backhed, 2013). Human gut microbes potentially benefit the host by synthesizing different vitamins, metabolizing bile acids, etc., thus exhibiting a fundamentally mutualistic association between some gut flora and the human host (Clarke et al., 2014). Therefore, microbes may be considered a supplemental "organ" in the host (Bäckhed et al., 2005). Furthermore, the number of microbial cells in the human body is reportedly approximately 10-fold the number of human cells (Rosner, 2014). Therefore, it is essential to systematically analyze associations between microbes and humans. The Human Microbiome Project (HMP) has furthered the current understanding of microbial structure, diversity, and function over the years (Human Microbiome Project Consortium, 2012). However, numerous basic and clinical studies have investigated the association between the human microbiome and human health (Moore and Moore, 1995; Dethlefsen et al., 2007; Zhang et al., 2009; Brown et al., 2011).

It is important to understand microbe-host interactions, which could benefit the prevention, diagnosis, treatment, and prognosis of human diseases (Bao et al., 2017; Zou et al., 2018). Microbial communities could be influenced by not only maternal genetic factors (Khachatryan et al., 2008; Turnbaugh et al., 2009; Goodrich et al., 2014) but also the habitat environments, such as the change of season (Davenport et al., 2014), host diet (David et al., 2014), antibiotic consumption (Donia et al., 2014), host smoking habits (Mason et al., 2015), and residential hygiene of the host (Sommer and Backhed, 2013). Changes in environmental variables may modify microbial communities and alter host-microbe interactions (Ma et al., 2014). In the past decades, with the development of high-throughput sequencing techniques and ensuing computational tools, increasing evidence demonstrates the close association between microbial dysbiosis and various human diseases (Neish, 2009), such as inflammatory bowel disease (IBD) (Frank et al., 2007), diabetes (Brown et al., 2011; Giongo et al., 2011), asthma (Chen and Blaser, 2007), obesity (Ley et al., 2006), and some cancers (Moore and Moore, 1995; Schwabe and Jobin, 2013). For example, through 16S rRNA microarray and parallel clone library-sequencing analysis, Huang et al. (2011) collected bronchial epithelial brushings from 65 asthma patients and compared them with 10 other samples from healthy control subjects, reporting that members of the airway microbiota, such as *Comamonadaceae*, *Sphingomonadaceae*, and *Oxalobacteraceae*, were greater in asthma patients. Hoppe et al. (2011) evaluated the effect of *Oxalobacter formigenes* on primary hyperoxaluria, a rare genetic disease. In particular, the urinary oxalate test and *ad hoc* analysis in their study revealed a reduction in *Oxalobacter formigenes* in patients with kidney stones. Furthermore, to analyze and elucidate the microbiota of colon cancer patients, Sobhani et al. (2011) extracted bacterial DNA from 179 colon cancer patients. Through qPCR and the immunohistochemical analyses, *C. coccoides*, *Bacteroides*, *Lactobacillus groups*, and *Faecalibacterium prausnitzii species* were reportedly increased in colon cancer patients. Moreover, on comparing microbes from 83 healthy control individuals and 98 liver cirrhosis patients, Qin et al. (2014) identified several biomarkers associated with liver cirrhosis, reporting that certain groups were reduced (e.g., *Alistipes finegoldii*, *Bacteroides eggerthii*, and *Coprococcus*) while certain others were enriched (e.g., *Fusobacterium*, *Haemophilus parainfluenzae*, and *Phascolarctobacterium*). Therefore, elucidation of the association between microbes and human diseases may facilitate novel drug discovery.

Despite some reported microbe-disease associations, they are not sufficient to completely understand disease pathogenesis, diagnosis, and treatment. Fortunately, Wang et al. (2015) proposed the excellent work about cancer hallmark network framework in the predictive genomics. The cancer hallmark network framework offered great insights on modeling genome sequencing data to predict cancer evolution and associated clinical phenotypes, which provided valuable designment strategies for using the framework in conjunction with genome sequencing data in any other attempt to prediction works on human diseases, drug targets and other fields, microbe included. Indeed, construction of a computationally efficient model from existing associations to predict potential ones is practical, potentially providing novel insights into time-consuming microbiology experiments by elucidating the most promising previously unknown associations (Chen et al., 2017c). Specifically, in determining lncRNA-disease associations (Chen, 2015), studies on drug targets (van Laarhoven et al., 2011; Yamanishi, 2013) and miRNA-disease associations (Wang et al., 2010; Chen et al., 2017d) have yielded various efficient *in silico* models to predict the underlying associations. Recently, based on experimentally verified microbe-disease associations, Ma et al. (2017) constructed the first Human Microbe-Disease Association Database (HMDAD). Thereafter, several computational models have been proposed to further contribute to the HMDAD. For example, Chen et al. (2017a) generated a model based on the KATZ measure, named KATZHMDA. In their model, they first constructed an association network showing pairwise relationships between microbes and human disease. Furthermore, they introduced Gaussian interaction profile kernel similarity for microbes and diseases to predict novel associations. Moreover, Huang Z.A. et al. (2017) developed a model of Path-Based Human Microbe-Disease Association Prediction (PBHMDA), wherein they used a special depth-first search algorithm on the heterogeneous biological network. In particular, they investigated all possible paths between diseases and microbes to infer highly probable associations. Resulting from the idea of collaborative recommendation model, Huang Y.A. et al. (2017) provided a computational model by adopting neighbor-based collaborative filtering and a graph-based scoring approach to calculate the association possibility of unknown microbe–disease pairs. The usage of hybrid approach based on two single recommendation methods contributed much more on their prediction results. Based on the microbe-disease interaction network, Wang et al. (2017) developed the model of Laplacian Regularized Least Squares for Human Microbe-Disease Association (LRLSHMDA). LRLSHMDA is a semi-supervised computational model using the Laplacian regularized least squares classifier.

Recently, Zou et al. (2018) integrated symptom-based disease similarity to predict novel human microbe-disease associations based on network consistency projection (NCPHMDA). In detail, they conducted microbe space projection and disease space projection and combined the projections to design an advisable non-parametric approach. Based on adaptive boosting approach, Peng et al. (2018) developed a model named Adaptive Boosting for Human Microbe-Disease Association prediction (ABHMDA) to reveal the underlying associations between microbes and human diseases by calculating the association probability of concerned disease-microbe pair by grouped weak classifiers to form a stronger classifier for further scoring and sorting samples. Not long time ago, Qu et al. (2019) proposed a computational model on the basis of HMDAD by the methods of matrix decomposition and label propagation, which divided the original adjacency matrix about the relationship between microbes and diseases into a linear combination of itself and a low-rank matrix to predict novel disease-microbe associations.

Herein, we present a Random Walk on Hypergraph for Microbe-Disease Association Prediction (RWHMDA) model to predict underlying microbe-disease associations. In particular, we constructed a higher-order hypergraph model to accurately determine the implicit inherent association between microbes and human diseases. Thereafter, we generalized the well-known random walk process to the hypergraph in a modified manner, wherein vertices (microbes) within a hyperedge (human disease) were differentiated by the walker depending on their features. Finally, we ranked all candidate microbes for every investigated human disease. The merit of this study is the introduction of the concept and method of hypergraph to predict microbe-disease associations. Hypergraph is practical and suitable because it could provide biologically decipherable aspects by placing all disease-associated microbes in one hyperedge. Furthermore, we implemented global and local Leave-one-out cross-validation (LOOCV) to evaluate the predictive performance of RWHMDA.

MATERIALS AND METHODS

Human Microbe-Disease Associations

In this study, we utilized microbe-disease associations in HMDAD database (Ma et al., 2017)[1], containing 483 known microbe-disease associations among 292 microbes inhabiting the human body and 39 human diseases. The associations in HMDAD were obtained from sequencing-based microbiological analyses. In addition, if different data are available for overlapping microbe-disease associations in the database, only one record would be maintained. Finally, we obtained 450 distinct known microbe-disease associations for further prediction. Microbe-disease associations could be stored in an adjacency matrix$A$, where element $A(i, j)$ represented the binary association of disease $d(i)$ and microbe $m(j)$. In other words, we obtained a $nd \times nm$ matrix A, where 450 elements were 1 and the others were 0. Meanwhile, nd was the number of diseases, and nm was the number of microbes.

[1]http://www.cuilab.cn/hmdad

Gaussian Interaction Profile Kernel Similarity for Microbes

Gaussian interaction profile kernel similarity was calculated on the basis of a type of Radial Basis Function (RBF), namely Gaussian kernel function. In this study, we adopted the Gaussian interaction profile kernel similarity to determine the similarity between microbes. In detail, based on the constructed adjacency matrix A, microbial interaction profiles could be defined as a binary vector IP($m(j)$), representing the absence or presence of the interaction between microbe $m(j)$ and diseases. IP($m(j)$) was the j-th column of matrix A. Thereafter, we calculated the Gaussian kernel similarity between microbe $m(j)$ and microbe $m(j)$, using Gaussian kernel function as follows:

$$GM\left(m\left(i\right),\ m\left(j\right)\right)\ =\ \exp\left(-r_m||IP\left(m\left(i\right)\right)-IP\left(m\left(j\right)\right)||^2\right) \tag{1}$$

where r_m was set to balance the kernel bandwidth, and GM defined the Gaussian interaction profile kernel similarity matrix for microbes. Specially, $r_m$was calculated in accordance with a new parameter r_m and the average known association number per miRNA as follows:

$$r_m = \frac{r'_m}{\left(\frac{1}{nm}\sum_{i=1}^{nm}||IP\left(m\left(i\right)\right)||^2\right)} \tag{2}$$

where nm is the total number of microbes. Technically, r_m was set as 1 here (Chen et al., 2017b).

RWHMDA

In this study, we proposed the RWHMDA model from the random walk on hypergraph to predict novel microbe-disease associations. Although Gaussian interaction profile kernel similarity for microbes is also accounted for in this method, RWHMDA is still a graph structure-based model without extra domain information in microbiological studies. Random walks on simple graphs have been investigated extensively in various biological fields. However, random walks on hypergraph have not been reported with respect to the prediction of microbe-disease associations thus far. Hypergraph is a type of higher-order graphical representation of biological data, compensating for information loss in the normal graph method, exclusively describing pair-wise association structures (**Figure 1**).

Generally, in the present model, we first constructed a hypergraph comprising microbes and diseases, wherein diseases are presented as hyperedges and microbes are presented as nodes. If several microbes have been confirmed to be associated with one disease, they would be presented as nodes in the hyperedge corresponding to the disease. In the hypergraph, hyperedges can join numerous vertices (not limited to two nodes as in simple graph). Specifically, if microbe $m(i)$ is associated with disease $d(j)$, then node $m(i)$ belongs to hyperedge $d(j)$. Obviously, one microbe might belong to different hyperedges. We assessed all known microbe-disease associations and established a hypergraph with 39 hyperedges. Without loss of generalizability,

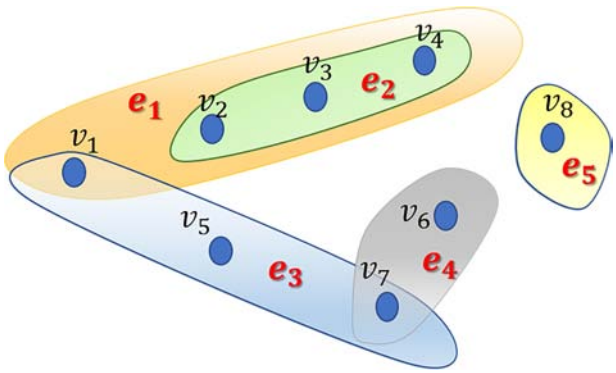


FIGURE 1 | An example of a hypergraph comprising 5 hyperedges and 8 nodes. Different hyperedges are indicated with different colors. Every hyperedge contains different numbers of nodes based on their practical applications.

we defined the hypergraph as $HG\ (V, \mathcal{E})$ where $\mathcal{E}$ is the set of hyperedges, and $_V$ is the set of vertices. Hyperedge $e \in \mathcal{E}$ is a subset of V, hyperedge e is incident with node v if the node belongs to the hyperedge. Neighborhood relationships among nodes can be defined if $v \in e, w \in e$.

After constructing the hypergraph, we implement random walk with restart on it. The random walk on the normal graph is a type of Markov process. The surfer travels between nodes in the graph by starting at a node and shifting to an adjacent node at each discrete time step t. The transition probability between nodes is completely independent of the time t. Therefore, we could define the transition probability matrix $P \in R^{|V|\times|V|}$ for the whole process. Matrix P represents the transition probabilities of the random internodal movements. Matrix P is actually a critical factor calculated on the basis of multifarious filed knowledge. Furthermore, we introduced the random walk on the hypergraph. Basically, the surfer shifts between two nodes only if they are neighbors in the currently visited hyperedge. Briefly, this process may be considered a two-step procedure as follows: the surfer randomly selects a hyperedge incident with a currently visited node in step 1; thereafter, the surfer selects a destination neighbor node within the selected hyperedge in step 2 (**Figure 2**). Thereafter, we would focus on capturing transition matrix P with respect to the random walk on the hypergraph.

Considering an unweighted hypergraph $HG\,(V, \mathcal{E})$, wherein hyperedges and nodes have no weights, the incidence matrix $H \in R^{|V|\times|E|}$ was defined as follows:

$$h\,(v, e) = \begin{cases} 1, & if\ v \in e \\ 0, & if\ v \notin e \end{cases} \tag{3}$$

$$\delta\,(e) = |e| \tag{4}$$

$$d\,(v) = |\mathcal{E}\,(v)| \tag{5}$$

where $\delta\,(e)$ is the degree of hyperedge e, $d(v)$ is the degree of vertex v, $|e|$ indicates the number of nodes within hyperedge e, $\mathcal{E}\,(v)$ is the set of hyperedges incident with vertex v. Thereafter, we obtained the diagonal hyperedge degree matrix $D_e \in R^{|E|\times|E|}$, the diagonal vertex degree matrix $D_v \in R^{|V|\times|V|}$.

Regarding data on microbe-disease associations, it means the surfer would select a disease known to be associated with the current microbe. We could not unambiguously distinguish the more critical disease associated with the referenced microbe. Therefore, we intend for the surfer to uniformly randomly select a hyperedge at step 1. Furthermore, the surfer would walk to a node within this hyperedge. In our predictive case, although it is potentially difficult to evaluate the features of nodes, we differentiated microbes within a hyperedge of disease in accordance with the Gaussian interaction profile kernel similarity. Technically, in step 2, we intend for the surfer to shift to a node within a hyperedge in accordance with the sum of similarities of the node with all other nodes in the hypergraph. In summary, starting from node u, the surfer would select hyperedge e incident with u proportional to the weight of hyperedge $w(e)$. Thereafter, the surfer selects node v proportional to the weight of v within the current hyperedge e, namely $w\,(v_e)$.

Considering the afore-mentioned motivation, we then defined the weighted incident matrix $W \in R^{|V|\times|E|}$ of hypergraph $HG(V, \mathcal{E})$ as follows:

$$w\,(v, e) = \begin{cases} w\,(v_e), & if\ v \in e \\ 0, & if\ v \notin e \end{cases} \tag{6}$$

where $w\,(v_e)$ is the weight of node v in hyperedge e. In the present model, we calculated $w\,(v_e)$ on the basis of matrix GM. In this study, the weight of a microbe $m(i)$ in a hyperedge is the sum of ith row in GM. Thereafter, we redefined hyperedge degree $\delta'\,(e)$ and hyperedge degree matrix $D_{ve} \in R^{|E|\times|E|}$ as follows:

$$\delta\,(e) = \sum_{v\in e} w\,(v, e) \tag{7}$$

where D_{ve} is the diagonal hyperedge degree matrix with element $\delta\,(e)$.

We then calculated the transition probability from vertex u to vertex v as follows:

$$P\,(u, v) = \sum_{e\in E} w\,(e)\, \frac{h\,(u, e)}{\sum_{\hat{e}\in\mathcal{E}} w\,(\hat{e})}\, \frac{w\,(v, e)}{\sum_{\hat{v}\in e} w\,(\hat{v}, e)} \tag{8}$$

which may also be expressed in matrix form as follows:

$$P = D_v^{-1} H W_e D_{ve}^{-1} W^T \tag{9}$$

where $W_e \in R^{|E|\times|E|}$ is the diagonal matrix of hyperedge weights, wherein all diseases are considered with equal weightage in accordance with the previously described practical consideration, i.e., all hyperedges are $1/|E|$. Naturally, transition matrix P is stochastic, implying that the sum of every row equals 1.

Furthermore, we implement the random walk with restart on the hypergraph. In particular, assuming that microbes associated with disease $d(i)$ are to be predicted, all microbes with known associations with $d(i)$ are considered seed microbes, while the others are considered candidate microbes. Thereafter, we set the initially normalized probability vector $\vec{v}\,(0) \in R^{|V|\times 1}$ such that seed microbes are assigned with equal probability and the

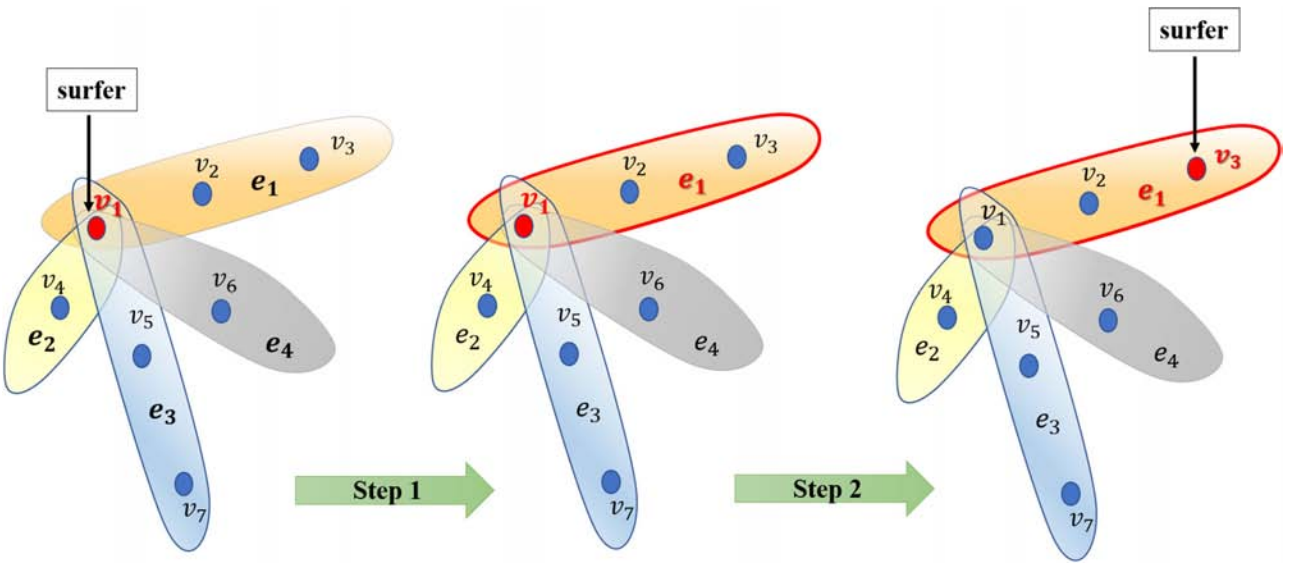


FIGURE 2 | Illustration of the random walk process on the hypergraph. Generally, the surfer selects a hyperedge in step 1 and then selects a node as the destination vertex to shift to the selected hyperedge in step 2.

non-seed miRNAs are zero. After the first step, $\vec{v}(1) = P^T\vec{v}(0)$. Moreover, we set the restart probability at every step at source nodes as α ($0 < \alpha < 1$), $\vec{v}(1) = (1-\alpha)P^T\vec{v}(0) + \alpha\vec{v}(0)$. Finally, we obtained the random walk with the following formula:

$$\vec{v}(t+1) = (1-\alpha)P^T\vec{v}(t) + \alpha\vec{v}(0) \quad (10)$$

$\vec{v}(t)$ is defined such that the ith element means the probability of moving to node i at step t. After some steps, the random walk would stabilize, implying that the difference between $\vec{v}(t+1)$ and $\vec{v}(t)$ measured by the L1 norm is smaller than the provided threshold. The stable state of the random walk with restart is defined as $\vec{v}(\infty)$. Stationary probability in $\vec{v}(\infty)$ indicates the probable associations between candidate microbes with the currently investigated disease. We conducted the random walk for every disease in the HMDAD database and ranked the underlying microbe-disease associations in accordance with the corresponding $\vec{v}(\infty)$ of the current disease (**Figure 3**).

As a supplement, we set the α-value as 0.2 and set the cutoff value as 10^{-6}.

RESULTS

Performance Evaluation

LOOCV was usually implemented to assess the performance of the prediction model. Global and local LOOCV in the present study were both conducted to comprehensively assess the performance of RWHMDA. Specifically, global LOOCV was conducted on the basis of the known microbe-disease associations in the HMDAD database (Ma et al., 2017). Each association was left out in turn as the test sample, while others were set as candidate samples. If the rank of the test sample was higher than that of the candidate samples, the test association was considered to have been correctly predicted. Furthermore, local LOOCV was somewhat different from global LOOCV, and it was implemented as follows: first, for an investigated disease, based on the association records in the HMDAD (Ma et al., 2017) database, each known disease-associated microbe was excluded in turn as the test sample and the others were used as seed samples. Thereafter, the predicted association probability of the current test sample would be ranked with the probability of candidate samples. If the test sample was ranked beyond the threshold, the model successfully predicted this microbe–disease association. Further, we plotted a receiver operating characteristics (ROC) curve. The area under the ROC curve (AUC) was determined to assess the prediction performance of RWHMDA. Specifically, AUC = 1 implied an excellent performance, and AUC = 0.5 indicated a random performance. Consequently, RWHMDA yielded a global AUC value of 0.8898 and local AUC value of 0.8524, which were higher than some previously reported computational models, such as LRLSHMDA (0.8959, 0.7657) (Bao et al., 2017) and KATZHMDA (0.8382, 0.6812) (Chen et al., 2017a; **Figure 4**).

Case Studies

To further assess the performance of the proposed model, we conducted case studies of asthma, Crohn's disease (CD), and type 2 diabetes by assessing the 10 highest probable microbes ranked by RWHMDA.

It is unambiguous that the human microflora play an important role in asthma pathogenesis (Li N. et al., 2017). Morbidity rates among asthma patients have significantly increased since the 1960s (Anandan et al., 2010). Asthma caused approximately 400 thousand deaths worldwide in 2015. More recently, on evaluating data regarding the association between *Helicobacter pylori* status with the history of asthma from 7663

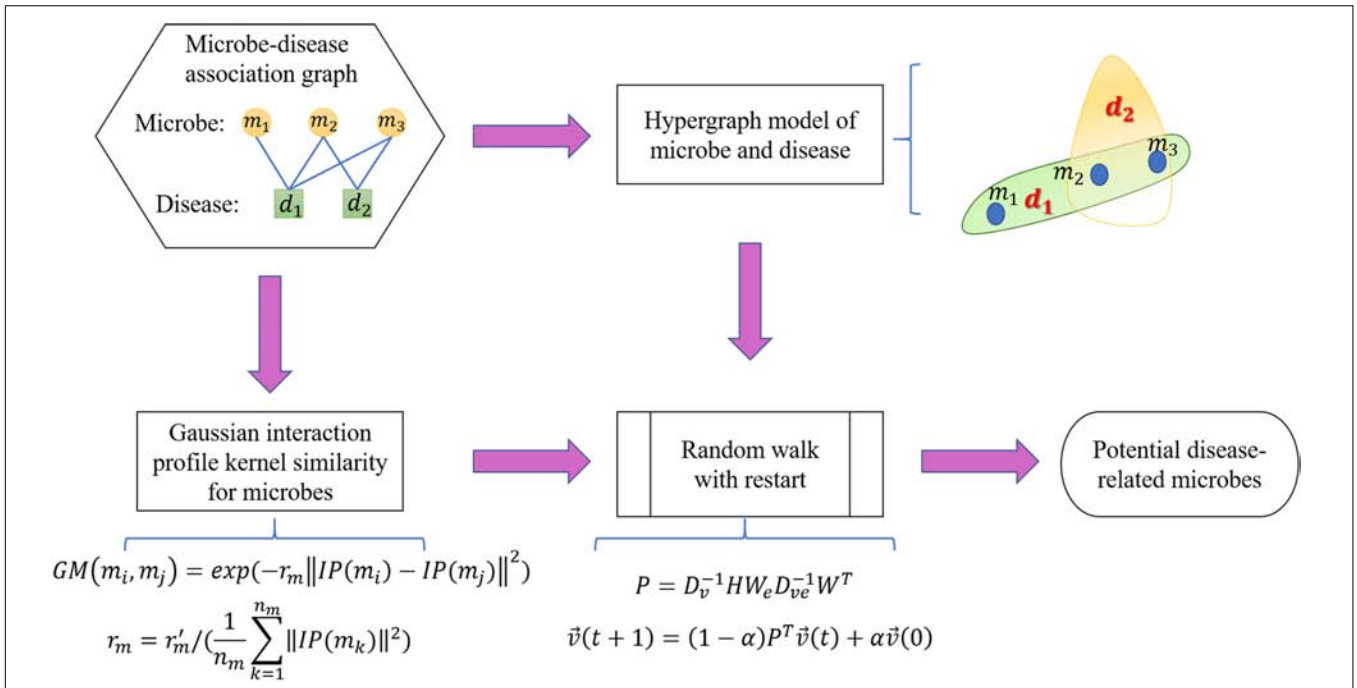


FIGURE 3 | Schematic representation of the RWHMDA model.

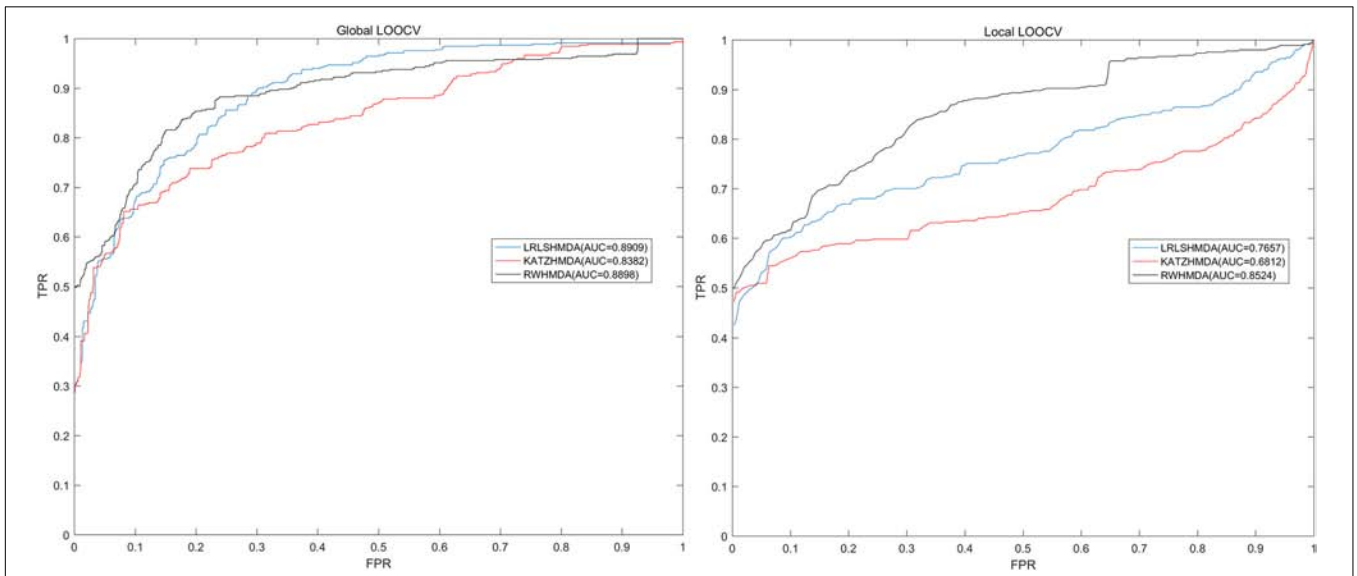


FIGURE 4 | Comparisons between the RWHMDA model and the other two state-of-the-art prediction models (LRLSHMDA and KATAHMDA) in terms of global and local AUC values. Consequently, RWHMDA yielded AUCs of 0.8898 and 0.8524, yielding a better prediction performance.

adults in the Third National Health and Nutrition Examination Survey, childhood acquisition of *H. pylori* is associated with a reduced risk of asthma (Chen and Blaser, 2007). We implemented RWHMDA for the asthma case study. Consequently, 9 of the 10 most highly ranked asthma-related microbes were confirmed from the literature (**Table 1**). For example, the study reporting the presence of *Propionibacterium acnes* (1st ranked in the prediction list of RWHMDA) in asthma patients helped diagnose asthma (Romero-Espinoza et al., 2018). *Pseudomonas*, ranked 3rd by our model, was confirmed to be more prevalent in the sputum of asthma patients (Jung et al., 2016). Moreover, as the 10th predicted asthma-related microbe, *Streptococcus* are associated with asthma, potentially contributing to its pathophysiology (Zhang et al., 2016).

The worldwide prevalence of diabetes mellitus has increased continuously over the past few decades

TABLE 1 | A case study on predicted potential asthma-related microbes.

Rank	Microbe	Evidence
1	*Propionibacterium acnes*	PMID:29447223
2	*Propionibacterium*	PMID:29447223
3	*Pseudomonas*	PMID:27433177
4	*Burkholderia*	PMID:24451910
5	*Enterobacter aerogenes*	PMID:18790035
6	*Enterobacter hormaechei*	Unconfirmed
7	*Klebsiella pneumoniae*	PMID:26220531
8	*Shigella dysenteriae*	Unconfirmed
9	*Actinobacteria*	PMID:23265859
10	*Streptococcus*	PMID:27078029

Among the 10 highest ranked potential asthma-related microbes, eight were confirmed from the literature.

(Tadic and Cuspidi, 2015). Type 2 diabetes mellitus is a subclass of diabetes mellitus, accounting for approximately 90% of all the diabetes mellitus cases. The traditional view holds that the pathogenesis of type 2 diabetes is associated with both genetic and lifestyle-related factors. Recent evidence suggests that the pathomechanism and pathogenesis of type 2 diabetes mellitus are also associated with an unbalance in microbial communities (Ripsin et al., 2009; Furet et al., 2010). Larsen et al. (2010) assessed the differences in the composition of the intestinal microbiota in individuals without and those with type 2 diabetes via high-throughput 16S rDNA gene pyrosequencing, reporting an increase in *Bacilli*, *Bacteroidetes*, and *Betaproteobacteria* and reductions in *Clostridia*, *Clostridium*, *Firmicutes*, etc. Among the 10 highest ranked microbes by probability, 8 were confirmed through recent evidence (**Table 2**). For example, *Fusobacterium nucleatum* was ranked first and confirmed to be significantly higher in type 2 diabetes mellitus patients than in those without type 2 diabetes mellitus (Miranda et al., 2017). *Pseudomonas*, abundant in the subgingival plaque, ranked 2nd by our model and was markedly different between individuals with and those without diabetes (Zhou et al., 2013). Furthermore, *Aerococcus* and *Atopobium* were associated with the risk of type 2 diabetes (Li H. et al., 2017; Long et al., 2017).

Crohn's disease (CD) is a type of IBD. Although the etiology of CD is generally believed to associated with the combination of immune, environmental, and bacterial factors, however, the precise etiology of CD is still unclear (Dessein et al., 2008; Stefanelli et al., 2008; Cho and Brant, 2011). In fact, no surgical treatment or pharmacotherapeutic methods have been reported to cure Crohn's disease (Baumgart and Sandborn, 2012). Studies have increasingly investigated the bacterial factors associated with the etiology of CD. Gevers et al. (2014) reported that the increased abundance of *Fusobacteriaceae*, *Enterobacteriaceae*, *Pasteurellacaea*, and *Veillonellaceae* and the decreased abundance of *Clostridiales Erysipelotrichales*, and *Bacteroidales* are closely correlated with Crohn's disease. A case study on Crohn's disease revealed that the 10 most probable microbes were confirmed through recent researches (**Table 3**). For example, the two most promising microbes predicted by our model were *Clostridium difficile* and *Bacteroides fragilis*, both confirmed to be present at high levels in CD patients compared than in healthy individuals (Cojocariu et al., 2014; Zhou et al., 2016). Moreover, studies evaluating the association between disease status and gut microbiota in CD patients revealed that *Clostridium coccoides* (3rd place in the ranking list) was abundant in febrile patients presenting with remission in comparison with patients with active CD (Prosberg et al., 2016).

TABLE 2 | RWHMDA used to predict candidate microbes associated with type 2 diabetes.

Rank	Microbe	Evidence
1	*Fusobacterium nucleatum*	PMID:28198980
2	*Pseudomonas*	PMID:23613868
3	*Aerococcus*	PMID:28786059
4	*Atopobium*	PMID:28177125
5	*Atopobium vaginae*	unconfirmed
6	*Candidate division TM7*	unconfirmed
7	*Eggerthella*	PMID:26046242
8	*Gardnerella*	PMID:28316574
9	*Gardnerella vaginalis*	PMID:2131794
10	*Lactobacillus crispatus*	PMID:28608654

Consequently, 8 of the 10 most probable microbes were experimentally confirmed through the relevant literature.

DISCUSSION AND CONCLUSION

Accumulating falsifiable evidence indicates that microbial involvement is associated with disease pathogenesis in some cases. In this study, with data from microbiological studies, hypergraph theory, and other research areas, we introduced an *in silico* model named RWHMDA to predict underlying microbe-disease associations. Many previous computational models performed pairwise comparisons and illustrated microbe-disease associations as a normal graph. RWHMDA has been developed primarily on the basis of a hypergraph, thus compensating for the information loss issue by normal graph. Known microbe-disease associations in the HMDAD database

TABLE 3 | A case study on Crohn's disease verifying all 10 of the 10 most probable candidates of Crohn's disease-related microbes.

Rank	Microbe	Evidence
1	*Clostridium difficile*	PMID:25599768
2	*Bacteroides fragilis*	PMID:27684872
3	*Clostridium coccoides*	PMID:27687331
4	*Bacilli*	PMID:29559804
5	*Betaproteobacteria*	PMID:27833911
6	*Lachnospiraceae*	PMID:26628508
7	*Clostridium*	PMID:29722832
8	*Prevotella*	PMID:28852861
9	*Alistipes finegoldii*	PMID:28877044
10	*Alistipes putredinis*	PMID:29311644

and the Gaussian interaction profile kernel similarity for microbes were utilized to design a weighted hypergraph comprising microbes and diseases. Random walk with restart was implemented on the hypergraph for every disease to identify the potential disease-associated microbes. Both cross-validation and case studies on asthma, type 2 diabetes, and Crohn's disease revealed the reliability of RWHMDA. In addition, the predicted microbes for all diseases were publicly released for further validation through biological assays (**Supplementary Table S1**).

Generally, RWHMDA performed reliably, thus revealing several important factors. First, as a representation of a higher-order structure, hypergraphs adequately illustrate and present data on microbe-disease associations without information loss. In particular, the practice of setting disease as a hyperedge and microbe as a node was reasonable and biologically decipherable, thereby naturally benefiting the prediction of potential associations. Second, owing to the valid and updated data on disease-microbe associations through numerous biological analyses, RWHMDA had a greater prediction accuracy with greater probability. Third, random walk process is a widespread and significant physical dynamic process, used extensively in numerous studies. RWHMDA was developed on the basis of the random walk with restart process, following a seemingly iterative 2-step walking strategy to investigate the potential association probability between any pair of microbe and disease.

However, the present RWHMDA model has some limitations. Hypergraph and Gaussian interaction profile kernel were both constructed largely on the basis of known associations. Therefore, the model may have a bias toward those well-known diseases and microbes. Furthermore, some other similarity measures of diseases could also be meticulously integrated into the RWHMDA model, such as symptom-based disease similarity and disease semantic similarity. Finally, the RWHMDA model could not be implemented for new diseases without known associations with microbes, being an inherent limitation of the graph-based model.

In conclusion, RWHMDA is expected to display promising potential to predict disease-microbe associations for follow-up experimental studies and facilitate the prevention, diagnosis, treatment, and prognosis of complex human diseases.

AUTHOR CONTRIBUTIONS

Y-WN developed the prediction method, conducted the experiments, analyzed the result, and wrote the manuscript. C-QQ, G-YY, and G-HW conceived the project, analyzed the result, and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

Y-WN, C-QQ, and G-HW were supported by the National Natural Science Foundation of China under Grant Nos. 11631014 and 11871311. G-YY was supported by the National Natural Science Foundation of China under Grant No. 11631014.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01578/full#supplementary-material

TABLE S1 | We prioritized candidate microbes for all the investigated human diseases in the HMDAD database. The prediction results for each disease were publicly released for further validation. The relatively high ranked disease-microbe associations were anticipated to be confirmed by biological experiments or future clinical observation.

REFERENCES

Anandan, C., Nurmatov, U., van Schayck, O. C., and Sheikh, A. (2010). Is the prevalence of asthma declining? Systematic review of epidemiological studies. *Allergy* 65, 152–167. doi: 10.1111/j.1398-9995.2009.02244.x

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science* 307:1915. doi: 10.1126/science.1104816

Bao, W., Jiang, Z., and Huang, D. S. (2017). Novel human microbe-disease association prediction using network consistency projection. *BMC Bioinform.* 18(Suppl. 16):543. doi: 10.1186/s12859-017-1968-2

Baumgart, D. C., and Sandborn, W. J. (2012). Crohn's disease. *Lancet* 380, 1590–1605. doi: 10.1016/s0140-6736(12)60026-9

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

Chen, X. (2015). KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5:16840. doi: 10.1038/srep16840

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2017a). 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

Chen, X., Niu, Y. W., Wang, G. H., and Yan, G. Y. (2017b). HAMDA: Hybrid approach for MiRNA-disease association prediction. *J. Biomed. Inform.* 76, 50–58. doi: 10.1016/j.jbi.2017.10.014

Chen, X., Niu, Y. W., Wang, G. H., and Yan, G. Y. (2017c). MKRMDA: multiple kernel learning-based Kronecker regularized least squares for MiRNA-disease association prediction. *J. Transl. Med.* 15:251. doi: 10.1186/s12967-017-1340-3

Chen, X., Wu, Q. F., and Yan, G. Y. (2017d). RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. *RNA Biol.* 14, 952–962. doi: 10.1080/15476286.2017.1312226

Chen, Y., and Blaser, M. J. (2007). Inverse associations of *Helicobacter pylori* with asthma and allergy. *Arch. Intern. Med.* 167, 821–827. doi: 10.1001/archinte.167.8.821

Cho, J. H., and Brant, S. R. (2011). Recent insights into the genetics of inflammatory bowel disease. *Gastroenterology* 140, 1704–1712. doi: 10.1053/j.gastro.2011.02.046

Clarke, G., Stilling, R. M., Kennedy, P. J., Stanton, C., Cryan, J. F., and Dinan, T. G. (2014). Minireview: gut microbiota: the neglected endocrine organ. *Mol. Endocrinol.* 28, 1221–1238. doi: 10.1210/me.2014-1108

Cojocariu, C., Stanciu, C., Stoica, O., Singeap, A. M., Sfarti, C., Girleanu, I., et al. (2014). Clostridium difficile infection and inflammatory bowel disease. *Turk. J. Gastroenterol.* 25, 603–610. doi: 10.5152/tjg.2014.14054

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

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

Dessein, R., Chamaillard, M., and Danese, S. (2008). Innate immunity in Crohn's disease: the reverse side of the medal. *J. Clin. Gastroenterol.* 42(Suppl. 3 Pt 1), S144–S147. doi: 10.1097/MCG.0b013e3181662c90

Dethlefsen, L., McFall-Ngai, M., and Relman, D. A. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 449, 811–818. doi: 10.1038/nature06245

Donia, M. S., Cimermancic, P., Schulze, C. J., Wieland Brown, L. C., Martin, J., Mitreva, M., et al. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* 158, 1402–1414. doi: 10.1016/j.cell.2014.08.032

Eckburg, P. B., Lepp, P. W., and Relman, D. A. (2003). Archaea and their potential role in human disease. *Infect. Immun.* 71, 591–596. doi: 10.1128/iai.71.2.591-596.2003

Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13780–13785. doi: 10.1073/pnas.0706625104

Furet, J. P., Kong, L. C., Tap, J., Poitou, C., Basdevant, A., Bouillot, J. L., et al. (2010). Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes* 59, 3049–3057. doi: 10.2337/db10-0253

Gevers, D., Kugathasan, S., Denson, L. A., Vazquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15, 382–392. doi: 10.1016/j.chom.2014.02.005

Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., Novelo, L. L., Casella, G., et al. (2011). Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J.* 5, 82–91. doi: 10.1038/ismej.2010.92

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

Hoppe, B., Groothoff, J. W., Hulton, S. A., Cochat, P., Niaudet, P., Kemper, M. J., et al. (2011). Efficacy and safety of *Oxalobacter formigenes* to reduce urinary oxalate in primary hyperoxaluria. *Nephrol. Dial. Transplant.* 26, 3609–3615. doi: 10.1093/ndt/gfr107

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

Huang, Z. A., Chen, X., Zhu, Z., Liu, H., Yan, G. Y., You, Z. H., et al. (2017). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Huang, Y. J., Nelson, C. E., Brodie, E. L., Desantis, T. Z., Baek, M. S., Liu, J., et al. (2011). Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *J. Allergy Clin. Immunol.* 127, 372.e3–381.e3. doi: 10.1016/j.jaci.2010.10.048

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

Jung, J. W., Choi, J. C., Shin, J. W., Kim, J. Y., Park, I. W., Choi, B. W., et al. (2016). Lung microbiome analysis in steroid-nasmall yi, ukrainianve asthma patients by using whole sputum. *Tuberc. Respir. Dis.* 79, 165–178. doi: 10.4046/trd.2016.79.3.165

Khachatryan, Z. A., Ktsoyan, Z. A., Manukyan, G. P., Kelly, D., Ghazaryan, K. A., and Aminov, R. I. (2008). Predominant role of host genetics in controlling the composition of gut microbiota. *PLoS One* 3:e3064. doi: 10.1371/journal.pone.0003064

Larsen, N., Vogensen, F. K., van den Berg, F. W., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., et al. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* 5:e9085. doi: 10.1371/journal.pone.0009085

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., Qi, T., Huang, Z. S., Ying, Y., Zhang, Y., Wang, B., et al. (2017). Relationship between gut microbiota and type 2 diabetic erectile dysfunction in Sprague-Dawley rats. *J. Huazhong Univ. Sci. Technol. Med. Sci.* 37, 523–530. doi: 10.1007/s11596-017-1767-z

Li, N., Qiu, R., Yang, Z., Li, J., Chung, K. F., Zhong, N., et al. (2017). Sputum microbiota in severe asthma patients: relationship to eosinophilic inflammation. *Respir. 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, J., Prince, A. L., Bader, D., Hu, M., Ganu, R., Baquero, K., et al. (2014). High-fat maternal diet during pregnancy persistently alters the offspring microbiome in a primate model. *Nat. Commun.* 5:3889. doi: 10.1038/ncomms4889

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe-disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Mason, M. R., Preshaw, P. M., Nagaraja, H. N., Dabdoub, S. M., Rahman, A., and Kumar, P. S. (2015). The subgingival microbiome of clinically healthy current and never smokers. *ISME J.* 9, 268–272. doi: 10.1038/ismej.2014.114

Miranda, T. S., Feres, M., Retamal-Valdes, B., Perez-Chaparro, P. J., Maciel, S. S., and Duarte, P. M. (2017). Influence of glycemic control on the levels of subgingival periodontal pathogens in patients with generalized chronic periodontitis and type 2 diabetes. *J. Appl. Oral Sci.* 25, 82–89. doi: 10.1590/1678-77572016-0302

Moore, W. E., and Moore, L. H. (1995). Intestinal floras of populations that have a high risk of colon cancer. *Appl. Environ. Microbiol.* 61, 3202–3207.

Neish, A. S. (2009). Microbes in gastrointestinal health and disease. *Gastroenterology* 136, 65–80. doi: 10.1053/j.gastro.2008.10.080

Peng, L. H., Yin, J., Zhou, L., Liu, M. X., and Zhao, Y. (2018). Human microbe-disease association prediction based on adaptive boosting. *Front. Microbiol.* 9:2440. doi: 10.3389/fmicb.2018.02440

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., et al. (2009). The NIH human microbiome project. *Genome Res.* 19, 2317–2323. doi: 10.1101/gr.096651.109

Prosberg, M., Bendtsen, F., Vind, I., Petersen, A. M., and Gluud, L. L. (2016). The association between the gut microbiota and the inflammatory bowel disease activity: a systematic review and meta-analysis. *Scand. J. Gastroenterol.* 51, 1407–1415. doi: 10.1080/00365521.2016.1216587

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

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

Ripsin, C. M., Kang, H., and Urban, R. J. (2009). Management of blood glucose in type 2 diabetes mellitus. *Am. Fam. Physician* 79, 29–36.

Romero-Espinoza, J. A., Moreno-Valencia, Y., Coronel-Tellez, R. H., Castillejos-Lopez, M., Hernandez, A., Dominguez, A., et al. (2018). Virome and bacteriome characterization of children with pneumonia and asthma in Mexico City during winter seasons 2014 and 2015. *PLoS One* 13:e0192878. doi: 10.1371/journal.pone.0192878

Rosner, J. L. (2014). Ten times more microbial cells than body cells in humans? *Microbe* 9:47. doi: 10.1128/microbe.9.47.2

Schwabe, R. F., and Jobin, C. (2013). The microbiome and cancer. *Nat. Rev. Cancer* 13, 800–812. doi: 10.1038/nrc3610

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

Sommer, F., and Backhed, F. (2013). The gut microbiota–masters of host development and physiology. *Nat. Rev. Microbiol.* 11, 227–238. doi: 10.1038/nrmicro2974

Stefanelli, T., Malesci, A., Repici, A., Vetrano, S., and Danese, S. (2008). New insights into inflammatory bowel disease pathophysiology: paving the way for novel therapeutic targets. *Curr. Drug Targets* 9, 413–418. doi: 10.2174/138945008784221170

Tadic, M., and Cuspidi, C. (2015). Type 2 diabetes mellitus and atrial fibrillation: from mechanisms to clinical practice. *Arch. Cardiovasc. Dis.* 108, 269–276. doi: 10.1016/j.acvd.2015.01.009

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, 480–484. doi: 10.1038/nature07540

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Ventura, M., O'Flaherty, S., Claesson, M. J., Turroni, F., Klaenhammer, T. R., van Sinderen, D., et al. (2009). Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat. Rev. Microbiol.* 7, 61–71. doi: 10.1038/nrmicro2047

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, E., Zaman, N., McGee, S., Milanese, J. S., Masoudi-Nejad, A., and O'Connor-McCourt, M. (2015). Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin. Cancer Biol.* 30, 4–12. doi: 10.1016/j.semcancer.2014.04.002

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Yamanishi, Y. (2013). Chemogenomic approaches to infer drug-target interaction networks. *Methods Mol. Biol.* 939, 97–113. doi: 10.1007/978-1-62703-107-3_9

Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2365–2370. doi: 10.1073/pnas.0812600106

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

Zhou, M., Rong, R., Munro, D., Zhu, C., Gao, X., Zhang, Q., et al. (2013). Investigation of the effect of type 2 diabetes mellitus on subgingival plaque microbiota by high-throughput 16S rDNA pyrosequencing. *PLoS One* 8:e61516. doi: 10.1371/journal.pone.0061516

Zhou, Y., Chen, H., He, H., Du, Y., Hu, J., Li, Y., et al. (2016). Increased *Enterococcus faecalis* infection is associated with clinically active Crohn disease. *Medicine* 95:e5019. doi: 10.1097/md.0000000000005019

Zou, S., Zhang, J., and Zhang, Z. (2018). Novel human microbe-disease associations inference based on network consistency projection. *Sci. Rep.* 8:8034. doi: 10.1038/s41598-018-26448-8

Conflict of Interest Statement: 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.

Copyright © 2019 Niu, Qu, Wang and Yan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 30 July 2019
doi: 10.3389/fmicb.2019.01723

Microbiota in Human Periodontal Abscess Revealed by 16S rDNA Sequencing

Jiazhen Chen[1†], Xingwen Wu[2†], Danting Zhu[3], Meng Xu[3], Youcheng Yu[2], Liying Yu[3] and Wenhong Zhang[1]**

[1] Department of Infectious Diseases, Huashan Hospital, Fudan University, Shanghai, China, [2] Department of Dentistry, Zhongshan Hospital, Fudan University, Shanghai, China, [3] Department of Dentistry, Huashan Hospital, Fudan University, Shanghai, China

Periodontal abscess is an oral infective disease caused by various kinds of bacteria. We aimed to characterize the microbiota composition of periodontal abscesses by metagenomic methods and compare it to that of the corresponding pocket and healthy gingival crevice to investigate the specific bacteria associated with this disease. Samples from abscess pus (AB), periodontal pocket coronally above the abscess (PO), and the gingival crevice of the periodontal healthy tooth were obtained from 20 periodontal abscess patients. Furthermore, healthy gingival crevice samples were obtained from 25 healthy individuals. Bacterial DNA was extracted and 16S rRNA gene fragments were sequenced to characterize the microbiota and determine taxonomic classification. The beta-diversity analysis results showed that the AB and PO groups had similar compositions. *Porphyromonas gingivalis*, *Prevotella intermedia*, and other *Prevotella* spp. were the predominant bacteria of human periodontal abscesses. The abundances of *Filifactor alocis* and *Atopobium rimae* were significantly higher in periodontal abscesses than in the periodontal pocket, suggesting their association with periodontal abscess formation. In conclusion, we characterized the microbiota in periodontal abscess and identified some species that are positively associated with this disease. This provides a better understanding of the components of periodontal abscesses, which will help facilitate the development of antibiotic therapy strategies.

Keywords: high-throughput sequencing, oral microbiota, periodontal abscess, 16S rDNA metagenomic, periodontal pocket

OPEN ACCESS

Edited by:
Qi Zhao,
Shenyang Aerospace University, China

Reviewed by:
Takuichi Sato,
Niigata University, Japan
Thuy Do,
University of Leeds, United Kingdom

****Correspondence:***
Liying Yu
wuyu1984@hotmail.com
Wenhong Zhang
zhangwenhong@fudan.edu.cn

[†]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: *09 March 2019*
Accepted: *12 July 2019*
Published: *30 July 2019*

Citation:
Chen J, Wu X, Zhu D, Xu M, Yu Y, Yu L and Zhang W (2019) Microbiota in Human Periodontal Abscess Revealed by 16S rDNA Sequencing. Front. Microbiol. 10:1723. doi: 10.3389/fmicb.2019.01723

INTRODUCTION

Periodontal abscess is an acute exacerbation of chronic periodontitis, exhibiting clinical symptoms of swelling and severe pain in the gingival margin. It is defined as a localized suppurative lesion that is related to periodontal alveolar bone loss and the accumulation of pus in the gingival wall of the periodontal pocket (Herrera et al., 2000). Previously, we cultured obligate anaerobic bacteria from periodontal abscess and characterized their antimicrobial resistance profiles (Xie et al., 2014), in which the predominant obligate anaerobes were black-pigmented *Prevotella*. Although the results were partly in agreement with the findings of previous studies (Jaramillo et al., 2005; Herrera et al., 2014), some bacteria such as those of the genus *Treponema* were unculturable and some

predominant anaerobes such as *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Fusobacterium* spp. were less frequently cultured, due to the culture condition or suitability of medium (Xie et al., 2014). In addition, it has been reported that the therapeutic effect of antibiotic regimens on periodontal abscess is limited (Smith and Davies, 1986; Herrera et al., 2000, 2014), suggesting the complexity of associated pathogens. Recently, we used high-throughput barcoded 16S rDNA sequencing to characterize the microbiota in the periodontal pocket of patients with periodontitis and compared these to those of patients with chronic obstructive pulmonary disease (COPD). As the number of different kinds of bacteria was determined in the subgingival plaque of every patient, we hypothesized that periodontal abscess is caused by a combination of microbiota, and specific pathogens might be more dominant in the abscess than in the pocket and in healthy controls suggesting a positive-association with abscess formation. Therefore, a clearer understanding of pathogens and the microbiota that cause periodontal abscess is necessary.

In this study, we used high-throughput barcoded 16S rDNA sequencing technique to characterize the microbiota of periodontal abscess, the corresponding pocket, and healthy gingival crevice to investigate the specific bacteria associated with periodontal abscess in human periodontitis.

MATERIALS AND METHODS

Patient Recruitment

Forty-five participants were recruited from March 2015 to September 2015, including 20 periodontitis patients with periodontal abscess and 25 periodontal healthy individuals. Subjects with the following conditions were excluded from the study: pregnancy, use of antibiotics or anti-inflammatory drugs during the past 3 months, and administration of periodontal therapy during the last 6 months. The study was approved by the ethics committee of Huashan Hospital, Fudan University (No. KY2014-023). All participants provided signed, informed consent. The study design is shown in **Supplementary Figure S1**. Probing depth (PD), clinical attachment loss (CAL), and simplified oral hygiene index (OHI-S) were assessed according to World Health Organization recommendations (WHO, 1997). Periodontitis was diagnosed as previously described (Wu et al., 2017) with the presence of more than one tooth with at least one site (mesiobuccal, buccal, distobuccal, mesiolingual, lingual, and distolingual sites) with PD $\geq$ 4 mm, CAL $\geq$ 2 mm, and bleeding on probing. Periodontal abscess was diagnosed by a periodontal specialist based on the patients' symptomatology, clinical and radiological examination findings such as swelling and enlargement of the gingiva, history of periodontal disease, and radiograph of the alveolar bone destruction around the cementoenamel junction. Patients with periodontal abscess but without periodontitis were excluded. Periodontal tooth health was defined as PD $\leq$ 2 mm and with no bleeding on probing at all six sites.

Specimen Collection and Isolation of Bacterial DNA

Three samples were collected from all the patients with abscess, including the sample of abscess pus (abscess group, AB), periodontal pocket coronally above the abscess (pocket group, PO), and gingival crevice of periodontal healthy tooth (patient control group, PC). Healthy teeth were also sampled from periodontally healthy individuals as the healthy control (control group, HC). The abscess samples were drained after decontamination of the mucosa. A No. 25# sterilized paper point (Gapadent, China) was immersed into the deep area of pus for 10 s after drainage with a sterilized probe. The periodontal pocket and healthy gingival crevice samples were dipped with No. 25# sterilized paper points as previously described (Wu et al., 2017). The periodontal pocket sample was not collected from one patient due to contamination with pus. All samples were stored in tris-EDTA buffer solution of pH 7.4 (Sigma, United States) in a freezer (−80°C). Bacterial DNA was extracted using the QiAamp DNA Mini Kit (Qiagen, Germany), according to the manufacturer's instructions.

Amplification of the 16S rDNA and Sequencing

According to previous studies (Claesson et al., 2010; Mizrahi-Man et al., 2013; Jorth et al., 2014), hypervariable V3–V4 or V4–V5 regions are recommended to study the microbiome when using the second-generation sequencing method. It was also determined that the V4–V5 region showed the best performance among all regions. Therefore, the amplification of the V4–V5 regions of 16S rDNA, library construction, index PCR, and PCR clean-up were performed as previously described (Wu et al., 2017). Equal amounts of tagged 16S rRNA gene amplicons of each sample were mixed and denatured with 0.1 M NaOH. The mixed library was diluted to a final concentration of 10–20 pM using 10 mM tris at pH 8.5. Multiplexed paired-end sequencing (2 $\times$ 300 bp reads) of the 16S rRNA amplicons was performed using a Miseq system (Illumina, San Diego, CA, United States). Image analysis and base calling were performed on the Miseq system using the MiSeq Reporter software (MSR). After de-multiplexing the data and removing the reads that failed the purity filter (PF = 0), the reads were converted to FASTQ format.

Data Analyses

The generated FASTQ files (.fastq) and quality files were acquired as raw and mapped sequence data using default settings of the QIIME2 software (version: 2018.8) (Caporaso et al., 2010). Each operational taxonomic unit (OTU) was generated with 97% similarity cutoff using UPARSE v7.1 and chimeric sequences were identified and removed using UCHIME. The phylogenetic affiliation of each 16S rRNA gene sequence was analyzed using RDP Classifier[1] based on the Silva (SSU132) 16S rRNA database using a confidence threshold of 70% (Amato et al., 2013; Duan X.B. et al., 2017; Duan X. et al., 2017; Xu et al., 2018). The output was based on the classification of reads at several

[1]http://rdp.cme.msu.edu/

taxonomic levels. The alpha- and beta-diversity analyses were computed from the previously constructed OTU table using Mothur software (v.1. 30.1) (Schloss et al., 2009) and weighted UniFrac (Lozupone et al., 2011) analysis. In the group level, abundance analysis was determined from rarefaction files by the Mann–Whitney test between patient and health control groups and a paired *t*-test in the self-control comparison (SPSS Statistics v20.0 and GraphPad Prism software v6.01). At the patient level, when analyzing the significantly dominant bacteria in abscess patients, any OTUs with abundance differences greater than 10% were considered significantly dominant. The minimum abundance cutoff was set at 0.1% abundance, and abundance values < 0.1% were neglected.

RESULTS

Participant Characteristics

Characteristics of the 45 enrolled subjects are listed in **Table 1**. The sex proportion and smoking status between patients and controls were not statistically different.

Taxonomic Classification of the 16S rDNA Sequences

In total, 4.1 GB raw data containing 1.70 million high-quality and classifiable reads were obtained from 84 samples. The sequencing depth was similar among the four groups as follows: 20.85 ± 3.16 k reads in the AB group, 15.50 ± 1.93 k reads in the PO group, 15.40 ± 1.40 k reads in the PC group, and 27.31 ± 7.21 k reads in the HC group. Among these high-quality reads, 99.41% were classified into 322 genera, belonging to 22 phyla, 38 classes, 84 orders, and 153 families. There was no significant difference in the proportion of unclassifiable sequences among the four groups based on a Kruskal–Wallis test ($P = 0.186$).

TABLE 1 | Characteristics data of the enrolled study participants.

	Patient (*N* = 20)	Health control (*N* = 25)
Gender (%)		
Male	11 (55.0%)	11 (44.0%)
Female	9 (45.0%)	14 (56.0%)
Age		
Mean (sd)	53.2 (16.2)	64.8 (6.7)
Smoking status (%)		
Non-smoker[a]	17 (85.0%)	20 (80.0%)
Smoker	3 (15.0%)	5 (20.0%)
Former smoker[b]	0	1 (4.0%)
Current smoker[c]	3 (15.0%)	4 (16.0%)
Cigarettes/day, Mean (SD)		
Smoker	12.3 (7.5)	16.4 (14.4)
Former smoker	0	20 (N/A)
Current smoker	12.3 (7.5)	15.5 (16.5)
Periodontal Index		
PD, Mean (SD)	4.8 (0.7)	2.5 (0.5)
CAL, Mean (SD)	5.8 (1.0)	2.8 (0.7)
OHI-S, Mean (SD)	2.43 (1.12)	1.45 (0.51)

[a]Non-smokers were those who either had never smoked or quit cigarettes at least 10 years prior to study entry. [b]Former smokers were those who quit cigarettes at least 6 months but < 10 years prior to study entry. [c]Current smokers were currently smokers or those who quit cigarettes < 6 months prior to study entry. PD, probing depth; CAL, clinical attachment loss; OHI-S, simplified oral hygiene index.

Alpha-Diversity Analysis

The alpha-diversity analysis was conducted with two indexes, namely the Shannon index implicating community diversity and the Chao1 index implicating community richness (**Supplementary Figure S2**). Plots were generated and exported to the rarefaction curves (Aagaard et al., 2012; **Supplementary Figure S3**). There was no significant difference in Chao1 and Shannon indexes among the four groups based on a one-way ANOVA analysis with Tukey's multiple comparisons test.

Beta-Diversity Analysis

The microbial raw OTU data were subjected to the principal coordinate analysis (PCoA) to evaluate the similarities among the four groups (**Figure 1**). The results showed that the samples from the AB and PO groups had similar microbiota compositions, which could be grouped into one cluster, whereas the HC group formed another cluster. The samples from the PC group could not be grouped into one cluster and were scattered in the 3D plot (**Figure 1**).

Abundance Analysis

Consistent with the results of the beta-diversity analysis, 8 of the top 10 most dominant bacteria were the same in these two groups. The genera *Porphyromonas*, *Treponema 2*, *Streptococcus*, *Neisseria*, *Fusobacterium*, *Prevotella*, *Prevotella 7* and *Tannerella* accounted for 61% of the bacteria in the AB and PO groups, and exhibited no significant differences between the AB and PO groups based on a paired *t*-test (**Table 3** and **Supplementary Figure S4**). Furthermore, based on the hierarchical clustering analysis of the four groups, 17 of the 20 (85%) AB samples and 11 of the 19 (58%) PO samples clustered together (lower cluster in **Figure 2**). In the PC and HC control groups, 8 of the top 10 most dominant bacteria overlapped. The genera *Streptococcus*, *Neisseria*, *Bacteroides*, *Fusobacterium*, *Veillonella*, *Prevotella*, *Actinomyces*, and *Porphyromonas* accounted for 57 and 59% of the total bacteria in the PC and HC groups, respectively (**Supplementary Figure S4**). Fourteen of the 20 (70%) PC samples and 22 of the 25 (88%) HC samples clustered together (upper cluster in **Figure 2**), indicating almost similar compositions between these two groups.

Dominant Bacteria in Abscesses at the Patient Level

In the AB group, the abundance of the most abundant OTUs in all samples did not exceed 50% (**Figure 3**). The data showed that 9 of 20 samples had two OTUs with an abundance over 10% and 7 of 20 samples had three or more OTUs with an abundance over 10%. Furthermore, 2 of 20 samples had two OTUs with an abundance over 20%, including the combination

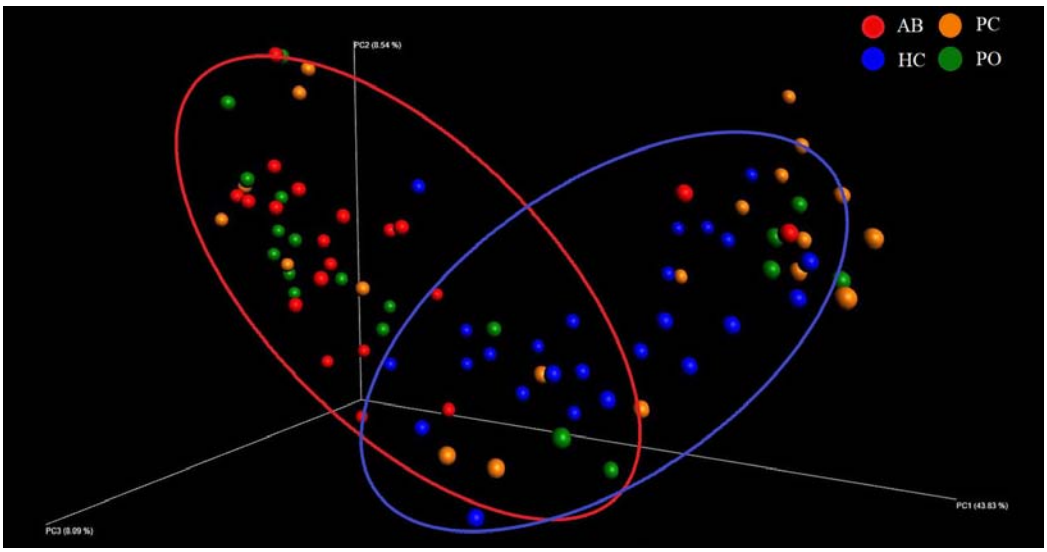


FIGURE 1 | Beta diversity analysis based on UniFrac analysis. Plots were generated using weighted UniFrac distances. Red dot represents the abscess pus (AB) group. Green dot represents the pocket (PO) group. Yellow dot represents the patient control (PC) group. Blue dot represents the healthy control (HC) group. Circles in red and blue represent different periodontal bacterial community clusters, respectively.

of *Leptotrichiaceae_Unclassified* (27.8%) and *P. gingivalis W83* (23.1%) in abscess 031711C, and *Lautropia_uncultured bacterium* (29.5%) and *Streptococcus_uncultured bacterium* (21.6%) in abscess 052004B (**Figure 3**). These results suggest that disease in these patients was caused by bacterial co-infections.

To reveal the dominant bacteria involved in abscess, we compared the OTUs of the AB group with those of the PC group by performing a paired *t*-test at the group level. The results showed that the abundance of 6 OTUs was significantly higher in the AB than in the PC group (**Supplementary Table S1**). However, except *P. gingivalis W83*, other classical opportunistic bacteria causing periodontal abscess were not significantly different based on this analysis. Furthermore, the OTUs, except *P. gingivalis W83*, which was relatively high in the AB group, were low in average abundance (**Supplementary Table S1**), suggesting that group comparison is not ideal for the analysis of dominant bacteria.

Considering the heterogeneity of dominant opportunistic bacteria in different patients, we performed a direct bacterial abundance comparison between the AB and PC groups at the patient level, and the bacteria with abundance differences > 10% between the AB and PC groups were identified as significantly dominant bacteria (**Table 2** and **Supplementary Table S2**). In total, 19 OTUs including *P. gingivalis W83* (8/20, 40%), *Prevotella* spp. (3/20, 15%), *Prevotella intermedia* (2/20, 10%), *P. gingivalis TDC60* (1/20, 5%), and *Prevotella heparinolytica* (1/20, 5%), were found to be significantly dominant in the AB group compared with abundances in the PC group in the corresponding number of patients (**Table 2**). Additionally, 20 OTUs, including *Streptococcus* spp. (6/20, 30%), *Actinomyces* spp. (3/20, 15%), *Lautropia* spp. (3/20, 15%), *Neisseria* spp. (3/20, 15%), *Veillonella* spp. (3/20, 15%), *Fusobacterium* spp. (2/20, 10%), *P. intermedia* (2/20, 10%), and *Bacteroides neonati* (2/20, 10%), were found to be significantly more dominant in the PC group than in AB group (**Table 2**). Interestingly, *P. intermedia* was identified as significantly dominant in the AB group of some patients and the PC group of other patients, indicating that it might have a heterogeneous function in abscess formation in different populations.

Specific Bacteria in Abscess Compared With Those in Pockets at the Group Level

At the group level, a comparison between the AB and PO groups revealed specific bacteria associated with acute disease. The abundance of bacteria in these two groups was highly similar. At the genus level, only *Filifactor* and *Atopobium* exhibited significantly higher abundance in the AB group, whereas nine genera presented significantly lower in abundance in the AB group than in the PO group (**Table 3**). Similarly, 3 OTUs including *Filifactor alocis* and *Atopobium rimae* presented significantly higher abundance in the AB group, suggesting that they might function in periodontal abscess formation. Moreover, 4 OTUs exhibited significantly lower abundance in the AB group than in the PO group (**Table 3**). However, all 4 OTUs were not accurately classified to the species level.

Bacteria Associated With Abscess at the Group Level

At the group level, a comparison between the AB and HC groups was made in the present study. At the genus level, 24 genera including *Porphyromonas*, *Treponema 2*, *Tannerella*, *Filifactor*, *Parvimonas*, and *Prevotella 1* were significantly more abundant in the AB group than in the HC group (**Supplementary Table S3**). Moreover, 25 genera including *Streptococcus*, *Neisseria*,

TABLE 2 | The Dominant OTUs in AB group compared to PC group at patient level.

OTUs	Patients (Radio)
Greater than 10% higher in AB group than in PC group	
Porphyromonas gingivalis W83	8 (40%)
Prevotella_uncultured bacterium	3 (15%)
Prevotella intermedia	2 (10%)
Streptococcus constellatus subsp. constellatus	2 (10%)
Fusobacterium_uncultured bacterium	2 (10%)
Porphyromonas_Unclassified	2 (10%)
Porphyromonas gingivalis TDC60	1 (5.0%)
Prevotella heparinolytica	1 (5.0%)
[Eubacterium] brachy	1 (5.0%)
Fusobacterium_uncultured organism	1 (5.0%)
Leptotrichiaceae_Unclassified	1 (5.0%)
Porphyromonas_uncultured organism	1 (5.0%)
Prevotella 1_Unclassified	1 (5.0%)
Rikenellaceae RC9 gut group_uncultured bacterium	1 (5.0%)
Streptococcus_uncultured bacterium	1 (5.0%)
Tannerella_Unclassified	1 (5.0%)
Treponema 2_Unclassified	1 (5.0%)
Treponema 2_uncultured Treponema sp.	1 (5.0%)
Treponema 2_uncultured bacterium	1 (5.0%)
Greater than 10% lower in AB group than in PC group	
Streptococcus_uncultured bacterium	6 (30%)
Actinomyces_uncultured bacterium	3 (15%)
Lautropia_uncultured bacterium	3 (15%)
Neisseria_Unclassified	3 (15%)
Veillonella_uncultured bacterium	3 (15%)
Prevotella intermedia	2 (10%)
Bacteroides neonati	2 (10%)
Fusobacterium_uncultured bacterium	2 (10%)
Porphyromonas gingivalis W83	1 (5.0%)
Bergeriella denitrificans	1 (5.0%)
Treponema 2_Unclassified	1 (5.0%)
Capnocytophaga_uncultured bacterium	1 (5.0%)
Corynebacterium_uncultured bacterium	1 (5.0%)
Fretibacterium_Unclassified	1 (5.0%)
Gemella_uncultured bacterium	1 (5.0%)
Haemophilus_uncultured bacterium	1 (5.0%)
Lentimicrobiaceae_uncultured bacterium	1 (5.0%)
Neisseria_uncultured bacterium	1 (5.0%)
Porphyromonas_uncultured bacterium	1 (5.0%)
Streptococcus_Unclassified	1 (5.0%)

The dominant different abundance of OTU was set as 10%.

Veillonella, *Capnocytophaga*, *Actinomyces*, *Selenomonas 3*, and *Prevotella 2* were significantly less abundant in the AB group than in the HC group (**Supplementary Table S3**). At the OTU level, 28 OTUs including *P. gingivalis*, *Treponema 2* spp., *P. intermedia*, *F. alocis*, and *T. forsythia* exhibited significantly higher abundance in the AB group than in the HC group. In contrast, 22 OTUs including *Streptococcus* spp., *Veillonella* spp., *Actinomyces* spp., and *Neisseria* spp. showed less abundance in the AB group than in the HC group (**Table 4**).

Porphyromonas gingivalis and *P. intermedia* were found to be dominant in the AB group at the patient level, and their abundance was significantly higher in the AB group than in the HC group at the group level. Additionally, *Prevotella* spp. was identified to be the dominant species in the abscesses of a few patients, but its abundance was not significantly different

TABLE 3 | Specific bacteria of significant difference in AB compared with PO group based on a paired *t*-test at group level.

Genus/OTUs	AB (Mean ± SE, %)	PO (Mean ± SE, %)	*p*-value
Higher in AB in Genus level			
Filifactor	2.05 ± 0.40	1.06 ± 0.26	0.009
Atopobium	0.75 ± 0.33	0.03 ± 0.01	0.046
Lower in AB in Genus level			
Leptotrichia	0.15 ± 0.05	0.96 ± 0.25	0.005
Selenomonas 3	0.12 ± 0.06	0.75 ± 0.29	0.029
Desulfobulbus	0.24 ± 0.13	0.61 ± 0.22	0.019
Prevotella 2	0.06 ± 0.02	0.52 ± 0.14	0.009
Flexilinea	0.09 ± 0.04	0.44 ± 0.16	0.048
Eikenella	0.02 ± 0.01	0.23 ± 0.09	0.037
F0332	0.03 ± 0.02	0.12 ± 0.05	0.029
Bergeyella	0.04 ± 0.02	0.11 ± 0.04	0.012
Lachnoanaerobaculum	0.02 ± 0.01	0.10 ± 0.03	0.013
Higher in AB in OTU level			
Streptococcus constellatus subsp. constellatus	2.08 ± 0.97	0.36 ± 0.17	0.049
Filifactor alocis ATCC 35896	1.88 ± 0.43	0.77 ± 0.24	0.007
Atopobium rimae	0.68 ± 0.31	0.00 ± 0.00	0.039
Lower in AB in OTU level			
Capnocytophaga_uncultured bacterium	0.26 ± 0.11	1.82 ± 0.67	0.027
Veillonella_uncultured bacterium	0.20 ± 0.09	1.42 ± 0.61	0.040
Desulfobulbus_Unclassified	0.24 ± 0.13	0.61 ± 0.22	0.019
Leptotrichia_uncultured bacterium	0.12 ± 0.04	0.55 ± 0.19	0.034

Abundance cutoff was set at 0.1%, some below 0.1% were not shown in the table.

between the AB and HC groups and it could not be classified at the species level. In contrast, *Streptococcus* spp., *Actinomyces* spp., *Lautropia* spp., *Neisseria* spp., and *Veillonella* spp. were the dominant species in the PC group compared with those in the AB group at the patient level, and their abundance was significantly lower in the AB group than in the HC group at the group level.

Specific Bacteria Associated With Periodontitis at the Group Level

We also compared the bacterial abundance between the PO and HC groups at the group level to determine if our data were consistent with well-known periodontitis-associated bacteria. At the genus level, 23 genera including genus *Porphyromonas*, *Treponema 2*, *Tannerella*, *Fretibacterium*, *Prevotella 1*, *Filifactor*, *Dialister*, and *Desulfobulbus* were significantly higher in the PO group than in the HC group, whereas 17 genera including *Streptococcus*, *Bacteroides*, *Veillonella*, *Bergeyella*, and *Kingella* were lower in the PO group than in the HC group (**Supplementary Table S4**). At the OTU level, the abundance of 29 OTUs including *P. gingivalis*, *P. intermedia*, *Treponema* spp., *T. forsythia*, *F. alocis*, and *P. heparinolytica* were higher in the PO group than in the HC group. The results were partly concordant with a previous study about the periodontal red and orange complex (Newman et al., 2015). In contrast, the abundance of 23 OTUs including *Streptococcus* spp. and *B. neonati* were lower in the PO group than in the HC group (**Supplementary Table S5**).

DISCUSSION

Oral periodontal abscess is an oral infective, painful disease that can spread (Yoneda et al., 2011; Herrera et al., 2014; Sato et al., 2016), and it is a valuable potential sign of undiagnosed type 2 diabetes (Alagl, 2017). Several studies have identified the dominant microbiota by culture-based diagnostic methods (Newman and Sims, 1979; Jaramillo et al., 2005; Xie et al., 2014). In the present study, considering the heterogeneity of dominant opportunistic bacteria in different patients, a patient level analysis between abscess and healthy periodontium was made, which showed that *P. gingivalis*, and *Prevotella* spp. including *P. intermedia* were found to be dominant in the abscess of some patients compared to those of healthy periodontium based on 16S rDNA metagenomic sequencing. Compared to the findings of our previous culture-based study, our study confirmed that *Prevotella* spp., and especially *P. intermedia*, is the dominant species in human periodontal abscess (Xie et al., 2014).

However, this study differs from the traditional culture-based method in the following aspects. First, the culture method usually detects the most abundant bacterium, but the second or the third most abundant bacteria can be neglected. For example, in abscess samples from two patients, *P. gingivalis W83* was the most dominant bacterium, and the abundance was 39.4 and

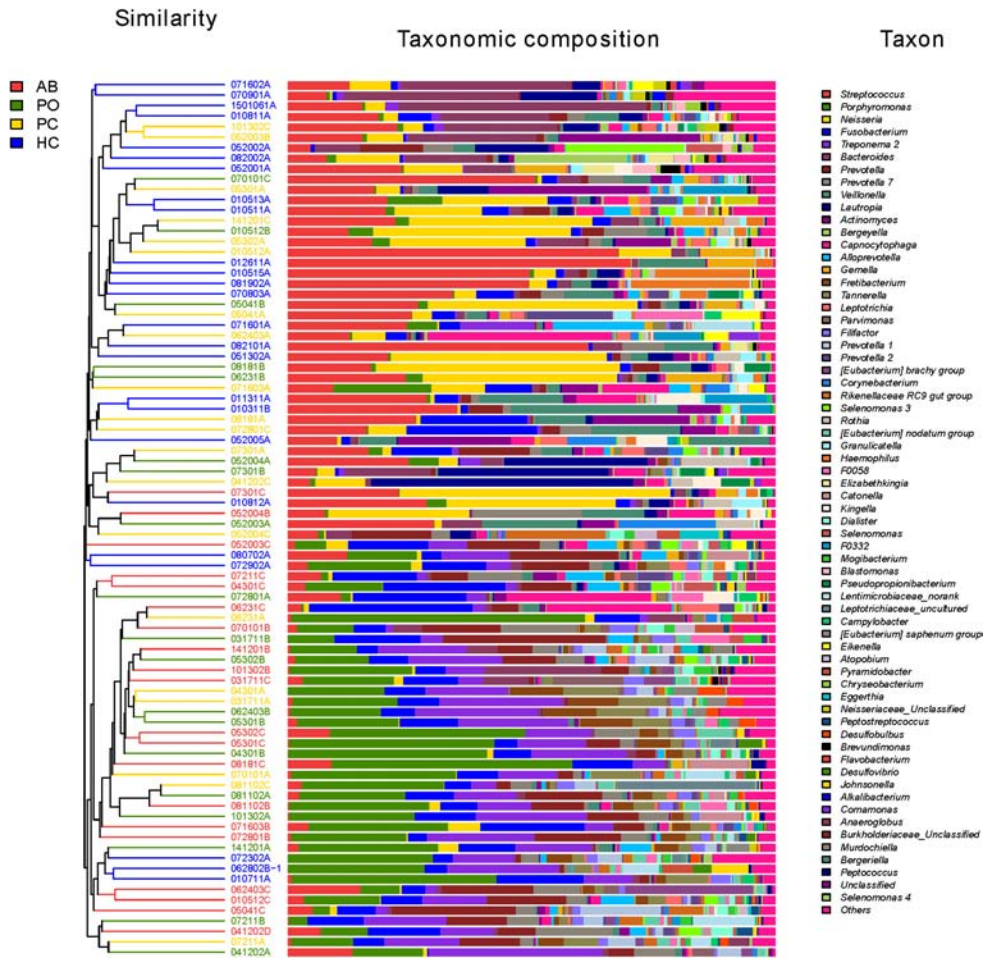


FIGURE 2 | Microbial community bar plot at the genus level with cluster tree. Each line shows the results of the bacterial community of one sample. Different bacterial taxonomies at the genus level (cutoff was set as 1%; some below 1% were shown as "others") are shown in different colors on the right. The hierarchical cluster diagrams in all samples based on the community composition (based on the algorithm of Bray–Curtis) is shown on the left. Red color indicates that the specimen is in the AB group. Green color indicates that the specimen is in the pocket (PO) group. Yellow color indicates that the specimen is in the healthy control (HC). Blue color indicates that the specimen is in the patient control (PC).

28.5%, respectively. Furthermore, the second highest abundant bacterium in both samples was *Fusobacterium* spp., for which abundance was only 12.3 and 17.9%, respectively, which could be neglected in the culture method. Second, the metagenomic sequencing method can detect unculturable bacteria and is not restricted to medium selectivity or addictive antibiotics. For example, in one abscess sample, the dominant bacterium was *Treponema 2* spp. (**Table 2**), which is unculturable. In addition, the major dominant bacterium identified in the present study was *P. gingivalis*, which was not detected in some by the culture method (Newman and Sims, 1979; Jaramillo et al., 2005; Xie et al., 2014). This might be attributed to medium selectivity or addition of selective antibiotics that inhibit this bacterium.

Porphyromonas gingivalis is a member of periodontal red complex (Socransky et al., 1998; Newman et al., 2015), which is the most predominant bacterial cluster detected in subgingival plaque, and can induce the production of interleukin-1 in macrophages (Saito et al., 1997) and trigger polyclonal B-cell activation (Champaiboon et al., 2000) associated with bleeding on probing and alveolar bone loss (Socransky et al., 1998). Moreover, it might be associated with several general dysfunctions including cardiovascular disease (Kozarov et al., 2005), rheumatoid arthritis (Berthelot and Le Goff, 2010), Alzheimer's disease (Dominy et al., 2019), and conception in women (Paju et al., 2017). *Prevotella* spp. including *P. intermedia* is a member of the periodontal orange complex (Newman et al., 2015), the second most predominant bacterial cluster detected in subgingival plaque, in addition to being recognized pathogens of periodontal infection. In a study by Jaramillo et al. (2005), the most frequent subgingival bacterium was *Fusobacterium* spp. (75%), followed by *P. intermedia* and *P. nigrescens* (60%), as well as *P. gingivalis* (51%). In partial agreement of the results of our previous study that *P. intermedia* is the most prevalent bacterium in periodontal abscess (Xie et al., 2014), in the present study, the second and third most dominant bacteria were *Prevotella* spp. and *P. intermedia* (25%).

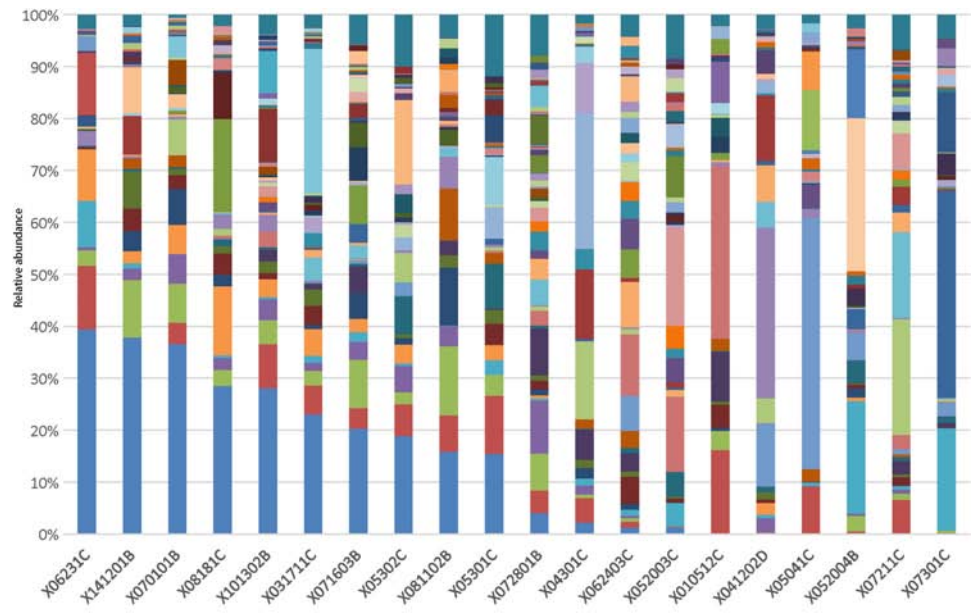


FIGURE 3 | Relative abundance of genera in the AB group. Each column represents a sample in the AB group. The genera were sorted in descending order of average abundance in abscess samples. Different bacterial taxonomies at the genus level are shown in different colors on the bottom. Abundance cutoff in this bar graph was set at 1%; some below 1% were shown as "others."

Like brain, lung, and pyogenic liver abscesses, which are caused by multiple kinds of bacteria (Brook, 2009; Webb et al., 2014; Yazbeck et al., 2014), periodontal abscess is more complex than previously thought. In the present study, seven of 20 samples had three or more OTUs with an abundance greater than 10%, and most OTUs were opportunistic bacteria, suggesting pathogen heterogeneity and bacterial co-infection in periodontal abscess diseases. Abscess occurs in a site that inhabits multiple normal and opportunistic bacteria, which are symbiotic and promote abscess formation (Newman et al., 2015). The significantly

TABLE 4 | Mean relative abundance of OTUs with significant statistical difference between AB and HC groups at group level.

OTUs	AB (Mean ± SE, %)	HC (Mean ± SE, %)	*p*-value
Higher in AB			
Porphyromonas gingivalis W83	13.64 ± 3.25	0.09 ± 0.04	< 0.001
Treponema 2_Unclassified	3.91 ± 0.86	0.75 ± 0.30	< 0.001
Prevotella intermedia	2.94 ± 1.31	0.17 ± 0.07	0.022
Treponema 2_uncultured bacterium	2.74 ± 1.62	0.46 ± 0.19	0.003
Porphyromonas_uncultured organism	2.63 ± 0.81	0.00 ± 0.00	< 0.001
Treponema 2_uncultured Treponema sp.	2.25 ± 0.58	0.08 ± 0.03	< 0.001
Streptococcus constellatus subsp. constellatus	1.98 ± 0.93	0.09 ± 0.05	0.006
[Eubacterium] brachy	1.95 ± 1.32	0.15 ± 0.07	0.018
Fretibacterium_Unclassified	1.90 ± 0.64	0.20 ± 0.09	0.004
Filifactor alocis ATCC 35896	1.79 ± 0.42	0.05 ± 0.03	< 0.001
Prevotella 1_Unclassified	1.77 ± 0.87	0.01 ± 0.01	0.003
Rikenellaceae RC9 gut group_uncultured bacterium	1.53 ± 0.99	0.00 ± 0.00	0.006
[Eubacterium] nodatum	1.37 ± 0.37	0.03 ± 0.02	< 0.001
Tannerella forsythia KS16	1.34 ± 0.52	0.05 ± 0.03	0.005
Prevotella heparinolytica	1.10 ± 0.83	0.00 ± 0.00	0.022
Eubacterium saphenum ATCC 49989	0.59 ± 0.27	0.02 ± 0.01	0.002
Parvimonas_uncultured bacterium	0.46 ± 0.23	0.02 ± 0.01	< 0.001
Defluviitaleaceae UCG-011_Unclassified	0.29 ± 0.14	0.03 ± 0.02	0.002
Rikenellaceae RC9 gut group_Unclassified	0.23 ± 0.07	0.00 ± 0.00	< 0.001
Clostridiales vadinBB60 group_Unclassified	0.22 ± 0.08	0.01 ± 0.01	0.003
Pseudoramibacter alactolyticus	0.19 ± 0.09	0.00 ± 0.00	0.016
Phocaeicola abscessus	0.18 ± 0.09	0.00 ± 0.00	0.001
Moryella_uncultured bacterium	0.15 ± 0.11	0.00 ± 0.00	0.002
Family XIII_uncultured bacterium	0.15 ± 0.05	0.01 ± 0.01	0.001
Mycoplasma hyosynoviae	0.12 ± 0.06	0.00 ± 0.00	0.006
W5053_uncultured bacterium	0.12 ± 0.07	0.00 ± 0.00	0.009
Flexilinea_Unclassified	0.11 ± 0.04	0.01 ± 0.00	0.003
Bulleidia_Unclassified	0.10 ± 0.09	0.00 ± 0.00	0.009
Lower in AB			
Streptococcus_uncultured bacterium	3.42 ± 1.40	13.21 ± 2.24	< 0.001
Streptococcus_Unclassified	0.56 ± 0.25	6.23 ± 1.78	< 0.001
Veillonella_uncultured bacterium	0.29 ± 0.12	4.35 ± 1.38	< 0.001
Actinomyces_uncultured bacterium	0.44 ± 0.10	3.20 ± 1.21	0.001
Neisseria_Unclassified	0.11 ± 0.06	2.98 ± 1.16	0.001
Capnocytophaga_uncultured bacterium	0.26 ± 0.11	2.94 ± 0.96	< 0.001
Prevotella 7_uncultured bacterium	0.79 ± 0.49	2.03 ± 0.60	0.005
Lautropia_uncultured bacterium	1.52 ± 1.47	1.68 ± 0.66	0.005
Gemella_uncultured bacterium	0.81 ± 0.58	1.63 ± 0.47	< 0.001
Leptotrichia_uncultured bacterium	0.12 ± 0.04	1.45 ± 0.34	< 0.001
F0332_uncultured bacterium	0.01 ± 0.01	1.13 ± 0.60	0.005
Granulicatella_uncultured bacterium	0.06 ± 0.02	0.81 ± 0.18	< 0.001
Selenomonas 3_uncultured bacterium	0.06 ± 0.05	0.81 ± 0.28	0.011
Blastomonas_uncultured bacterium	0.00 ± 0.00	0.52 ± 0.23	0.006
Ruminococcaceae UCG-014_uncultured bacterium	0.00 ± 0.00	0.45 ± 0.23	0.006
Kingella_uncultured bacterium	0.00 ± 0.00	0.44 ± 0.20	< 0.001
Delftia_Unclassified	0.00 ± 0.00	0.29 ± 0.13	0.011
Flavobacterium_Unclassified	0.04 ± 0.02	0.25 ± 0.08	0.011
Brevundimonas_uncultured bacterium	0.01 ± 0.01	0.24 ± 0.07	0.002
Aggregatibacter_uncultured bacterium	0.00 ± 0.00	0.22 ± 0.11	0.012
Ralstonia_Unclassified	0.00 ± 0.00	0.18 ± 0.06	0.012
Lachnoanaerobaculum_uncultured bacterium	0.01 ± 0.00	0.14 ± 0.05	0.009

Abundance cutoff was set at 0.1%, some below 0.1% were not shown in the table.

dominant bacteria in the abscess were also diverse in different patients and the difference between the abscess and the pocket remains unknown.

To the best of our knowledge, the bacteria in the periodontal abscess and periodontal pocket were compared for the first time. Periodontal abscess might represent acute exacerbation of periodontitis that is favored by changes in the subgingival microbiota, with an increase in bacterial virulence or a decrease in host defense (Herrera et al., 2014), resulting in the disruption of chronic phase (pocket) homeostasis and conversion to the acute phase (abscess). It is noteworthy that only two OTUs, *F. alocis* and *A. rimae*, were significantly higher in abundance in the AB group than in the PO group at the group level, indicating that they could be associated with the exacerbation of chronic periodontitis to acute periodontal abscess, although bacterial abundance and diversity were highly similar between the AB and PO groups. *F. alocis* is a Gram-positive anaerobic rod, which is now suggested to be a new periodontal pathogen (Schlafer et al., 2010; Aruni et al., 2015; Camelo-Castillo et al., 2015) with unique properties such as resistance to oxidative stress (Aruni et al., 2011), the ability to cause chronic inflammation (Fine et al., 2013), and the capacity to trigger apoptosis of gingival epithelial cells (Moffatt et al., 2011). In the present study, consistent with the findings of previous studies that *F. alocis* is positively associated with periodontitis, *F. alocis* was found to be more abundant in periodontal pockets than in the healthy periodontium. Furthermore, it was more abundant in periodontal abscess than in the pocket, suggesting that it is a potential, acute abscess-related, periodontal pathogen. *A. rimae* is an anaerobic, Gram-positive, rod-shaped bacterium, which has been suggested to be an endodontic abscess-related microorganism (Tennert et al., 2014; George et al., 2016). In the present study, *A. rimae* was found to be more significantly abundant in the abscess than in the pocket and healthy periodontium of the same patient. However, it has been reported that *A. rimae* is more prevalent in healthy subjects (Kumar et al., 2003), suggesting its role in periodontitis formation, which is complex and requires further study.

The bacteria associated with periodontitis have been well investigated previously (Liu et al., 2012; Wang et al., 2013). In the present study, the finding that the abundance of 29 OTUs including *P. gingivalis*, *P. intermedia*, *T. forsythia*, and *F. alocis* was higher in the PO group than in the HC group was largely in agreement with the findings of previous studies (Liu et al., 2012; Wang et al., 2013). These data further strengthen the reliability of this study to investigate the opportunistic pathogens and dominant microbiota associated with periodontal abscess.

Meanwhile, there were some limitations to our study. First, only the V4–V5 region, and not the full-length gene, was sequenced, which might result in some unclassified OTUs like *Streptococcus_Unclassified* at the species level. Second, different bacterial databases could lead to differences in detected species, which requires further comparisons with previously published studies to confirm the suitability of the present research. Third, the present study did not quantify the bacterial loads in samples, and quantitative research will be more helpful in unraveling the relationship between the severity of periodontal abscess and certain bacteria.

In conclusion, we used 16S rRNA-based metagenomics to characterize the bacterial profile of periodontal abscess in humans and compared it with the corresponding periodontal pocket and healthy periodontium. The results showed that the bacterial composition of periodontal abscess is more complex and mainly involves bacterial co-infections. Further, *P. gingivalis*, *P. intermedia*, and *Prevotella* spp. were the predominant bacteria in human periodontal abscesses. Two species, *F. alocis* and *A. rimae*, were found to be positively associated with abscess formation, although their bacterial abundance and diversity in periodontal abscess and periodontal pockets were highly similar. Recognition of the bacterial profile of periodontal abscess might reveal new strategies for the diagnosis, surveillance, and treatment of periodontal abscess, including the accurate use of antibiotics and probiotics.

DATA AVAILABILITY

All sequencing data were uploaded and deposited to the SRA database with project number PRJNA547446.

ETHICS STATEMENT

This study was approved by the Ethics Committee of the Huashan Hospital, Fudan University (No. KY2014-023).

AUTHOR CONTRIBUTIONS

JC and XW conceived and designed the study, acquired, analyzed, and interpreted the data, and drafted and critically revised the manuscript. DZ, MX, and YY acquired the data and critically revised the manuscript. LY and WZ conceived and designed the study, analyzed and interpreted the data, and drafted and critically revised the manuscript. All authors approved the final manuscript and agreed to be accountable for all aspects of the work.

FUNDING

This work was supported by the National Natural Science Foundation of China (81471987).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01723/full#supplementary-material

REFERENCES

Aagaard, K., Riehle, K., Ma, J., Segata, N., Mistretta, T. A., Coarfa, C., et al. (2012). A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One* 7:e36466. doi: 10.1371/journal.pone.0036466

Alagl, A. S. (2017). Periodontal abscess as a possible oral clinical sign in the diagnosis of undiagnosed diabetes mellitus of elderly in a dental clinic set up - a 7-year cross-sectional study. *J. Investig. Clin. Dent.* 8:e12217. doi: 10.1111/jicd.12217

Amato, K. R., Yeoman, C. J., Kent, A., Righini, N., Carbonero, F., Estrada, A., et al. (2013). Habitat degradation impacts black howler monkey (*Alouatta pigra*) gastrointestinal microbiomes. *ISME J.* 7, 1344–1353. doi: 10.1038/ismej.2013.16

Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., and Fletcher, H. M. (2015). Filifactor alocis–a new emerging periodontal pathogen. *Microb. Infect.* 17, 517–530. doi: 10.1016/j.micinf.2015.03.011

Aruni, A. W., Roy, F., and Fletcher, H. (2011). Filifactor alocis has virulence attributes that can enhance its persistence under oxidative stress conditions and mediate invasion of epithelial cells by *Porphyromonas gingivalis*. *Infect. Immun.* 79, 3872–3886. doi: 10.1128/IAI.05631-11

Berthelot, J. M., and Le Goff, B. (2010). Rheumatoid arthritis and periodontal disease. *Joint Bone Spine* 77, 537–541. doi: 10.1016/j.jbspin.2010.04.015

Brook, I. (2009). Microbiology and antimicrobial treatment of orbital and intracranial complications of sinusitis in children and their management. *Int. J. Pediatr. Otorhinolaryngol.* 73, 1183–1186. doi: 10.1016/j.ijporl.2009.01.020

Camelo-Castillo, A., Novoa, L., Balsa-Castro, C., Blanco, J., Mira, A., and Tomas, I. (2015). Relationship between periodontitis-associated subgingival microbiota and clinical inflammation by 16S pyrosequencing. *J. Clin. Periodontol.* 42, 1074–1082. doi: 10.1111/jcpe.12470

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, 335–336. doi: 10.1038/nmeth.f.303

Champaiboon, C., Yongvanitchit, K., Pichyangkul, S., and Mahanonda, R. (2000). The immune modulation of B-cell responses by *Porphyromonas ginginvalis* and interleukin-10. *J. Periodontol.* 71, 468–475. doi: 10.1902/jop.2000.71.3.468

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., et al. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res.* 38:e200. doi: 10.1093/nar/gkq873

Dominy, S. S., Lynch, C., Ermini, F., Benedyk, M., Marczyk, A., Konradi, A., et al. (2019). *Porphyromonas gingivalis* in Alzheimer's disease brains: evidence for disease causation and treatment with small-molecule inhibitors. *Sci. Adv.* 5:eaau3333. doi: 10.1126/sciadv.aau3333

Duan, X.-B., Wu, T.-X., Guo, Y.-C., Zhou, X.-D., Lei, Y.-L., Xu, X., et al. (2017). Marginal bone loss around non-submerged implants is associated with salivary microbiome during bone healing. *Int. J. Oral Sci.* 9, 95–103. doi: 10.1038/ijos.2017.18

Duan, X., Wu, T., Xu, X., Chen, D., Mo, A., Lei, Y., et al. (2017). Smoking may lead to marginal bone loss around non-submerged implants during bone healing by altering salivary microbiome: a prospective study. *J. Periodontol.* 88, 1297–1308. doi: 10.1902/jop.2017.160808

Fine, D. H., Markowitz, K., Fairlie, K., Tischio-Bereski, D., Ferrendiz, J., Furgang, D., et al. (2013). A consortium of *Aggregatibacter actinomycetemcomitans* (Aa), *Streptococcus parasanguinis* and Filifactor alocis are present in sites prior to bone loss in a longitudinal study of localized aggressive periodontitis. *J. Clin. Microbiol.* 51, 2850–2861. doi: 10.1128/jcm.00729-13

George, N., Flamiatos, E., Kawasaki, K., Kim, N., Carriere, C., Phan, B., et al. (2016). Oral microbiota species in acute apical endodontic abscesses. *J. Oral Microbiol.* 8:30989. doi: 10.3402/jom.v8.30989

Herrera, D., Alonso, B., de Arriba, L., Santa Cruz, I., Serrano, C., and Sanz, M. (2014). Acute periodontal lesions. *Periodontol* 65, 149–177. doi: 10.1111/prd.12022

Herrera, D., Roldan, S., and Sanz, M. (2000). The periodontal abscess: a review. *J. Clin. Periodontol.* 27, 377–386. doi: 10.1034/j.1600-051x.2000.027006377.x

Jaramillo, A., Arce, R. M., Herrera, D., Betancourth, M., Botero, J. E., and Contreras, A. (2005). Clinical and microbiological characterization of periodontal abscesses. *J. Clin. Periodontol.* 32, 1213–1218. doi: 10.1111/j.1600-051X.2005.00839.x

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., and Whiteley, M. (2014). Metatranscriptomics of the human oral microbiome during health and disease. *mBio* 5, e01012–e01014. doi: 10.1128/mBio.01012-14

Kozarov, E. V., Dorn, B. R., Shelburne, C. E., Dunn, W. A. Jr., and Progulske-Fox, A. (2005). Human atherosclerotic plaque contains viable invasive *Actinobacillus actinomycetemcomitans* and *Porphyromonas gingivalis*. *Arterioscler. Thromb. Vasc. Biol.* 25, e17–e18. doi: 10.1161/01.ATV.0000155018.67835.1a

Kumar, P., Griffen, A., Barton, J., Paster, B., Moeschberger, M., and Leys, E. (2003). New bacterial species associated with chronic periodontitis. *J. Dent. Res.* 82, 338–344. doi: 10.1177/154405910308200503

Liu, B., Faller, L. L., Klitgord, N., Mazumdar, V., Ghodsi, M., Sommer, D. D., et al. (2012). Deep sequencing of the oral microbiome reveals signatures of periodontal disease. *PLoS One* 7:e37919. doi: 10.1371/journal.pone.0037919

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172. doi: 10.1038/ismej.2010.133

Mizrahi-Man, O., Davenport, E. R., and Gilad, Y. (2013). Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS One* 8:e53608. doi: 10.1371/journal.pone.0053608

Moffatt, C. E., Whitmore, S. E., Griffen, A. L., Leys, E. J., and Lamont, R. J. (2011). Filifactor alocis interactions with gingival epithelial cells. *Mol. Oral Microbiol.* 26, 365–373. doi: 10.1111/j.2041-1014.2011.00624.x

Newman, M. G., and Sims, T. N. (1979). The predominant cultivable microbiota of the periodontal abscess. *J. Periodontol.* 50, 350–354. doi: 10.1902/jop.1979.50.7.350

Newman, M. G., Takei, H. H., and Klokkevold, P. R. (2015). *Carranza's Clinical Periodontology*, 12th Edn. St. Louis, MI: Elsevier.

Paju, S., Oittinen, J., Haapala, H., Asikainen, S., Paavonen, J., and Pussinen, P. J. (2017). Porphyromonas gingivalis may interfere with conception in women. *J. Oral Microbiol.* 9:1330644. doi: 10.1080/20002297.2017.1330644

Saito, A., Sojar, H. T., and Genco, R. J. (1997). Interleukin-1 gene expression in macrophages induced by surface protein components of *Porphyromonas gingivalis*: role of tyrosine kinases in signal transduction. *Oral Microbiol. Immunol.* 12, 135–140. doi: 10.1111/j.1399-302x.1997.tb00369.x

Sato, J., Kuroshima, T., Wada, M., Satoh, A., Watanabe, S., Okamoto, S., et al. (2016). Use of FDG-PET to detect a chronic odontogenic infection as a possible source of the brain abscess. *Odontology* 104, 239–243. doi: 10.1007/s10266-015-0218-211

Schlafer, S., Riep, B., Griffen, A. L., Petrich, A., Hubner, J., Berning, M., et al. (2010). Filifactor alocis–involvement in periodontal biofilms. *BMC Microbiol.* 10:66. doi: 10.1186/1471-2180-10-66

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

Smith, R., and Davies, R. (1986). Acute lateral periodontal abscesses. *Br. Dent. J.* 161, 176–178. doi: 10.1038/sj.bdj.4805920

Socransky, S. S., Haffajee, A. D., Cugini, M. A., Smith, C., and Kent, R. L. (1998). Microbial complexes in subgingival plaque. *J. Clin. Periodontol.* 25, 134–144. doi: 10.1111/j.1600-051x.1998.tb02419.x

Tennert, C., Fuhrmann, M., Wittmer, A., Karygianni, L., Altenburger, M. J., Pelz, K., et al. (2014). New bacterial composition in primary and persistent/secondary endodontic infections with respect to clinical and radiographic findings. *J. Endod.* 40, 670–677. doi: 10.1016/j.joen.2013.10.005

Wang, J., Qi, J., Zhao, H., He, S., Zhang, Y., Wei, S., et al. (2013). Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Sci. Rep.* 3:1843. doi: 10.1038/srep01843

Webb, G. J., Chapman, T. P., Cadman, P. J., and Gorard, D. A. (2014). Pyogenic liver abscess. *Frontline Gastroenterol.* 5, 60–67. doi: 10.1136/flgastro-2013-100371

WHO (1997). *Oral Health Surveys: Basic Methods*, 4th Edn. Geneva: World Health Organization.

Wu, X., Chen, J., Xu, M., Zhu, D., Wang, X., Chen, Y., et al. (2017). 16S rDNA analysis of periodontal plaque in chronic obstructive pulmonary disease and periodontitis patients. *J. Oral Microbiol.* 9:1324725. doi: 10.1080/20002297.2017.1324725

Xie, Y., Chen, J., He, J., Miao, X., Xu, M., Wu, X., et al. (2014). Antimicrobial resistance and prevalence of resistance genes of obligate anaerobes isolated from periodontal abscesses. *J. Periodontol.* 85, 327–334. doi: 10.1902/jop.2013.130081

Xu, L., Chen, X., Wang, Y., Jiang, W., Wang, S., Ling, Z., et al. (2018). Dynamic alterations in salivary microbiota related to dental caries and age in preschool children with deciduous dentition: a 2-Year follow-up study. *Front. Physiol.* 9:342. doi: 10.3389/fphys.2018.00342

Yazbeck, M. F., Dahdel, M., Kalra, A., Browne, A. S., and Pratter, M. R. (2014). Lung abscess: update on microbiology and management. *Am. J. Ther.* 21, 217–221. doi: 10.1097/MJT.0b013e3182383c9b

Yoneda, M., Kato, S., Mawatari, H., Kirikoshi, H., Imajo, K., Fujita, K., et al. (2011). Liver abscess caused by periodontal bacterial infection with *Fusobacterium necrophorum*. *Hepatol. Res.* 41, 194–196. doi: 10.1111/j.1872-034X.2010.00748.x

Conflict of Interest Statement: 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.

Copyright © 2019 Chen, Wu, Zhu, Xu, Yu, Yu and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 13 September 2019
doi: 10.3389/fmicb.2019.02120

Incorporating Statistical Test and Machine Intelligence Into Strain Typing of *Staphylococcus haemolyticus* Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry

Chia-Ru Chung[1†], *Hsin-Yao Wang*[2,3†], *Frank Lien*[2], *Yi-Ju Tseng*[2,4,5], *Chun-Hsien Chen*[2,4], *Tzong-Yi Lee*[6,7], *Tsui-Ping Liu*[2], *Jorng-Tzong Horng*[1,8*] and *Jang-Jih Lu*[2,9,10*]

[1] Department of Computer Science and Information Engineering, National Central University, Taoyuan City, Taiwan, [2] Department of Laboratory Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan City, Taiwan, [3] Ph.D. Program in Biomedical Engineering, Chang Gung University, Taoyuan City, Taiwan, [4] Department of Information Management, Chang Gung University, Taoyuan City, Taiwan, [5] Research Center for Emerging Viral Infections, Chang Gung University, Taoyuan City, Taiwan, [6] School of Life and Health Sciences, The Chinese University of Hong Kong, Shenzhen, China, [7] Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China, [8] Department of Bioinformatics and Medical Engineering, Asia University, Taoyuan City, Taiwan, [9] College of Medicine, Chang Gung University, Taoyuan City, Taiwan, [10] Department of Medical Biotechnology and Laboratory Science, Chang Gung University, Taoyuan City, Taiwan

Staphylococcus haemolyticus is one of the most significant coagulase-negative staphylococci, and it often causes severe infections. Rapid strain typing of pathogenic *S. haemolyticus* is indispensable in modern public health infectious disease control, facilitating the identification of the origin of infections to prevent further infectious outbreak. Rapid identification enables the effective control of pathogenic infections, which is tremendously beneficial to critically ill patients. However, the existing strain typing methods, such as multi-locus sequencing, are of relatively high cost and comparatively time-consuming. A practical method for the rapid strain typing of pathogens, suitable for routine use in clinics and hospitals, is still not available. Matrix-assisted laser desorption ionization-time of flight mass spectrometry combined with machine learning approaches is a promising method to carry out rapid strain typing. In this study, we developed a statistical test-based method to determine the reference spectrum when dealing with alignment of mass spectra datasets, and constructed machine learning-based classifiers for categorizing different strains of *S. haemolyticus*. The area under the receiver operating characteristic curve and accuracy of multi-class predictions were 0.848 and 0.866, respectively. Additionally, we employed a variety of statistical tests and feature-selection strategies to identify the discriminative peaks that can substantially contribute to strain typing. This study not only incorporates statistical test-based methods to manage the alignment of mass spectra datasets but also provides a practical means to accomplish rapid strain typing of *S. haemolyticus*.

Keywords: *Staphylococcus haemolyticus*, strain typing, MALDI-TOF MS, Fisher's exact test, machine learning

OPEN ACCESS

Edited by:
Qi Zhao,
Shenyang Aerospace University, China

Reviewed by:
Martin Welker,
Biomerieux, France
Ewa Szczuka,
Adam Mickiewicz University in Poznań, Poland

***Correspondence:**
Jorng-Tzong Horng
horng@db.csie.ncu.edu.tw
Jang-Jih Lu
janglu45@gmail.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: *01 June 2019*
Accepted: *28 August 2019*
Published: *13 September 2019*

Citation:
Chung C-R, Wang H-Y, Lien F, Tseng Y-J, Chen C-H, Lee T-Y, Liu T-P, Horng J-T and Lu J-J (2019) Incorporating Statistical Test and Machine Intelligence Into Strain Typing of Staphylococcus haemolyticus Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. Front. Microbiol. 10:2120. doi: 10.3389/fmicb.2019.02120

INTRODUCTION

Staphylococcus haemolyticus is one of the most significant species among the coagulase-negative staphylococci (CoNS), whose main ecological niches are skin and the human and animal mucous membranes (Becker et al., 2014). They are often the causative agents of septicemia, peritonitis, otitis, and urinary tract infections. In particular, the multidrug resistance, the early acquisition of resistance to methicillin, and various glycopeptide antibiotics by this species has troubled patients for many years (Froggatt et al., 1989; Hiramatsu, 1998). Strain typing of pathogenic *S. haemolyticus* forms an important part of the response to modern public health infectious disease outbreaks (MacCannell, 2013). For example, an outbreak of *S. haemolyticus* had been reported to be the cause of burn wound infections after a serious explosion event in Taiwan during June 2015 (van Duin et al., 2016; Chang et al., 2018). Rapid typing of *S. haemolyticus* facilitates the identification of the origin of infection, and allows rapid infection control when patients are critically ill. Consequently, a cost effective and rapid identification strategy that targets strain typing issues is essential and needs to be incorporated in routine clinical microbiology laboratory practices.

Whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) is widely used in clinical microbiology laboratories worldwide. This is because MALDI-TOF MS allows rapid, reliable, and cost-effective identification of bacterial species (Vrioni et al., 2018; Wang et al., 2018c). The MALDI-TOF mass spectrum contains extensive information regarding the matter that constitutes microorganisms. In addition to the identification of bacterial species, MALDI-TOF MS has the potential to allow strain typing and/or antibiotic resistance profiling with high accuracy when machine learning methods are also implemented (Croxatto et al., 2012; Mather et al., 2016). Compared to the other strain typing methods, such as pulse-field gel electrophoresis and multi-locus sequence typing (MLST), analysis by MALDI-TOF MS to determine strain type is advantageous owing to its lower cost and rapid turn-around-time (Wang et al., 2018b). Strain typing via MALDI-TOF MS is promising; however, the subtle differences in MALDI-TOF MS spectra of different strains has hindered the introduction of this type of analysis in a clinical context in the absence of incorporation of computational methods (Sandrin et al., 2013; Camoez et al., 2016). Numerous methods have been developed in recent years to overcome this drawback in strain typing by MALDI-TOF spectrum analysis. The visual examination of a MALDI-TOF pseudo-gel or spectrum to pinpoint strain-specific peaks has been implemented by some research groups (Wolters et al., 2011; Josten et al., 2013). Visual examination of the MALDI-TOF MS is easy in practice, but the analytical accuracy is highly dependent on the operator. Inter-batch and/or intra-batch analytical variation is extremely likely. Moreover, visual examination of a MALDI-TOF MS or pseudo-gel is labor-intensive. Analyzing complex proteomic data, such as those obtained by MALDI-TOF MS, by visual examination often does not attain the appropriate level of precision, adequate objectivity, and/or a high enough throughput.

With the rapid advancements in artificial intelligence, machine learning-based methods have been implemented to identify classifiers when facing such classification problems (Mather et al., 2016; Wang et al., 2018b). More specifically, the logistic regression (LR), support vector machine (SVM), the decision tree (DT), the random forest (RF), and k-nearest neighbor (KNN) approaches have been widely implemented to build classifier model systems. In recent years, the application of machine learning-based methods in the field of medicine has received considerable attention, and several studies have demonstrated that the use of artificial intelligence to analyze complex data in medical practice is apposite and promising (Shameer et al., 2018; Hannun et al., 2019). Specifically, machine learning-based classifiers allowing professional diagnosis of retinopathy (Gulshan et al., 2016), can be used to analyze electrocardiography data (Hannun et al., 2019), and have been used to predict the prognoses of diseases (Wang et al., 2016; Yu et al., 2016; Lin et al., 2018). In addition to image analysis, applying machine learning-based methods to proteomic studies, specifically MALDI-TOF MS investigations, has assisted in attaining high accuracy in strain type prediction and/or strain antibiotic resistance (Wang et al., 2018a,b,c). Machine learning-based methods are able to utilize the signal intensities of specific peaks in their predictions, and this provides additional and more improved information than those obtained by the traditional method based on the presence or absence of peaks (Walker et al., 2002; Wolters et al., 2011; Lasch et al., 2014). In addition to providing robust prediction accuracy, machine learning-based methods, when analyzing MALDI-TOF MS, are also able to generate sets of discriminative peaks that are essential for accurate prediction. These specific sets of discriminative peaks can be used to pinpoint the possible combinations of molecules that are responsible for the various strain types and the variation in drug resistance profiles (Vrioni et al., 2018).

As mentioned previously, slight differences in MALDI-TOF MS results among different strains should be considered critical in preprocessing the spectral data. Specifically, the determination or extraction of representative features is essential before constructing the classifiers. Yet, little research is being done to develop a definitive strategy to solve such issues, not to mention incorporating statistical tests. In this study, we first developed a statistical test-based strategy for dealing with the alignment issue for the MALDI-TOF MS according to the mass-to-charge ratio (m/z) values, and further considered the signal intensity to construct the classification models. Various machine learning algorithms were trained and validated with the aim of discriminating the ST3, ST42, and various other STs of *S. haemolyticus*. We also investigated the discriminative peaks that are central to strain typing of *S. haemolyticus* with MALDI-TOF MS. This approach will not only be beneficial in rapid outbreak control for *S. haemolyticus* infection but also provide a definite strategy for preprocessing the spectral data.

MATERIALS AND METHODS

Bacterial Isolates

A total of 254 unique *S. haemolyticus* isolates had been collected at Chang Gung Memorial Hospital, Linkou branch, Taiwan. The

period of collection was between June and November 2015, which was the period when a significant number of burn patients were admitted to the hospital. The isolates were stored at −70°C until use. This was a retrospective study investigating the relation between MS spectrum and microbial strain typing. No diagnosis or treatment was involved by the study. Waiver of informed consent was approved by the Institutional Review Board of Chang Gung Medical Foundation (No. 201600049B0).

Analytical Measurement of MALDI-TOF MS

To carry out the analysis, we cultivated the isolates on blood agar plates (Becton Dickinson, MD, USA) initially in a batch manner. The isolates were cultured in 5% CO_2 incubator for 16 h. We then conducted the analytical measurements required for MALDI-TOF MS following manufacturer's instructions. First, we picked a single colony from a blood agar plate and spread it onto a steel target plate as a thin film (Bruker Daltonik GmbH, Bremen, Germany). One μl of 70% formic acid (Bruker Daltonik GmbH, Bremen, Germany) was then applied onto the steel target plate followed by drying in room air. One μl of matrix solution (Bruker Daltonik GmbH, Bremen, Germany) was then added. After the sample preprocessing, a MicroFlex LT mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany) using a linear positive model was used for data acquisition. For each batch, a Bruker Daltonics Bacterial test standard (Bruker Daltonik GmbH, Bremen, Germany) was analyzed to allow calibration. The sampling setting of the laser shot was 240 shots (20 Hz) for each isolate. The MALDI-TOF MS spectra were analyzed using Biotyper 3.1 software (Bruker Daltonik GmbH, Bremen, Germany). The analytical range of each spectrum was 2,000-20,000 m/z. *S. haemolyticus* identification was set at high confidence (score > 2 in the reports of Biotyper 3.1 software). Furthermore, FlexAnalysis 3.3 (Bruker Daltonik GmbH, Bremen, Germany) was also implemented to acquire the numerical spectra data which derived from MALDI-TOF MS. Specifically, the original signals were smoothed by Savitzky-Golay algorithm and their baselines were subtracted by the top hat method. Meanwhile, some thresholds that were adopted to extract reasonable peaks were setup as explained below: signal-to-noise ratio was 2, relative intensity and minimum intensity were both 0, maximal number of peaks was 200, peak width was 6, and height was 80%. On the basis of the single measurements, we hypothesized that strain typing of *S. haemolyticus* is possible when the variability issue is handled using information engineering technology.

Multilocus Sequence Typing of *S. haemolyticus*

We defined the strain typing of *S. haemolyticus* by sequencing seven housekeeping genes, namely *arc*, *SH1200*, *hemH*, *leuB*, *SH1431*, *cfxE*, and *RiboseABC* (Panda et al., 2016). The sequencing results of these genes were used to assign the sequence types of *S. haemolyticus* throughout the present analysis using the MLST database (https://pubmlst.org/shaemolyticus/) powered by the BIGSdb genomics platform (Jolley et al., 2018).

MS Data Preprocessing for Classifiers Construction

Several computational tools have been developed for the preprocessing and extraction of features from MS data (Wong et al., 2005; Mantini et al., 2007; Gibb and Strimmer, 2012). More specifically, spectral data preprocessing would transform a set of raw spectra into a numerical table which include mass-to-charge (m/z) states with associated intensity for each isolate. Generally, m/z values with adequate intensities are considered as the fingerprint signatures when using spectral data, and these can be extracted to build up models for discriminating different subgroups. Note that a peak has an m/z value. As a result, a valuable analysis would highly depend on the appropriate use of preprocessing techniques. The MS data derived from FlexAnalysis 3.3 were of high quality, but their resulting peaks were not aligned within the dataset. Meanwhile, the aforementioned tools lack of specific information about the reference spectrum when implementing the alignment of peaks. Therefore, we developed a statistical test-based method for determining the reference spectrum within a given dataset, then further realizing the alignment of the peaks.

The reference spectrum should be capable of discriminating between different subgroups within a dataset. Consequently, we mainly focused on determining what pattern of peaks in the reference spectrum can indicate the differences among different groups in this study. For each spectrum, we first rounded each m/z value to the nearest whole number, and then all peaks that occurred were used to form a set of named candidate peaks set (CPS). The peaks in CPS were then sorted into ascending order. After a tolerance value is suggested, each adjacent peak in the CPS is either lower than or is equal to the given tolerance value; in such circumstances, the one with the higher difference in occurring ratio is retained. The difference in occurring ratio for m/z = k, in Dalton (Da), is defined below.

$$D_k = \frac{1}{3}\{|\frac{x_1}{n_1} - \frac{x_2}{n_2}| + |\frac{x_1}{n_1} - \frac{x_3}{n_3}| + |\frac{x_2}{n_2} - \frac{x_3}{n_3}|\},$$

where x_1, x_2, and x_3 are the counts that are aligned to m/z = k, and n_1, n_2, and n_3 are the number of isolates for ST3, ST42, and other ST types, respectively. For example, suppose that the tolerance value is 1 Da, and the CPS = (2428 Da, 2429 Da, 2435 Da, 2436 Da, 2437 Da, 2450 Da) with $D_{2428} = 0.053$, $D_{2429} = 0.090$, $D_{2435} = 0.080$, $D_{2436} = 0.094$, $D_{2437} = 0.076$, and $D_{2450} = 0.120$, then the final m/z values, 2429 Da, 2436 Da, and 2450 Da, are then used to create the representative peaks set (RPS), which has an ascending order. In other words, the RPS is the reference spectrum and feature set used to construct the classification models.

To analyze the common peaks across the datasets given in this study, we employed Fisher's exact test (Raymond and Rousset, 1995) to determine a tolerance value for constructing the RPS due to relatively small sample sizes. For each tolerance value, there are three p-values determined by comparing ST3 and ST42, ST3 and other ST types, and ST42 and other ST types. As mentioned previously, the reference spectrum should

be capable of discriminating between different subgroups within a dataset, and the tolerance value could be adopted according to its ability of separating these three groups. Therefore, the tolerance value was selected based on the obtained reference spectrum that would produce the largest number of *p*-values that were less than 0.001. We then further adopted the repeated 5-fold cross validation to demonstrate the efficiency of the determined tolerance value. Note that the determination of CPS and RPS was based on the training data when the repeated 5-fold cross validation was used. In other words, the repeated 5-fold cross validation was implemented here to simulate an external validation for evaluation of the performance in the determination of the reference spectrum. The flowchart of preprocessing is shown in **Figure 1**.

After determining the RPS, the alignment of the m/z with intensity is another critical part of the process, whereby the strength of signal at a specific m/z is determined. Therefore, in these circumstances, it is straightforward to move the specific m/z value of an isolate to the closest one in the RPS. As the tolerance value increases, more than one m/z values might be aligned to the same specific m/z in the RPS. In this situation, the intensity with the minimum distance between its own m/z and the specific m/z, is preserved. Hence, duplication problems can be solved. For instance, if both m/z = 2530 Da and m/z = 2535 Da in a spectrum are aligned to 2532 Da, which is a member of the RPS, the intensity of the m/z = 2530 Da is used for representing the strength of signal at 2532 Da. **Supplementary Figure 1** illustrates how this alignment takes place.

Development of Machine Learning-Based Classifiers

In this study, we implemented four machine learning methods; multiple logistic regression (MLR), support vector machine (SVM) learning, decision tree (DT) learning, and random forest (RF) learning, to construct the strain type classifiers for *S. haemolyticus* using R software (version 3.5.1, R Foundation for Statistical Computing, https://www.r-project.org/). MLR is a basic parametric model used in dealing with the present types of classification problems. The primary objective of SVM is to find a hyperplane that is able to segregate different classes of data and therefore it is commonly used to solve classification problems. DT and RF are both non-parametric tree-based strategies. Owing to the small size of data, the unsophisticated structure of DT can help us interpret the important features of the data more clearly. On the other hand, RF can provide evaluation metrics for the features and thus is able to identify the important features used during the model construction.

The glmnet package (Friedman et al., 2010) of R was applied during this study to construct the MLR model. More specifically, the MLR model can be defined as

$$P(G = k|X = x) = \frac{\exp(\beta_{0k} + \beta_k^T x)}{\sum_{j=1}^{K} \exp(\beta_{0j} + \beta_j^T x)}$$

where K is the number of levels of the response variable, and $G = (1, 2, \ldots, K)$ is the set of levels. Note that this parameterization is not estimable due to identical probabilities. However, regularization is able to deal with this. Hence the MLR model can be obtained by maximizing the penalized log-likelihood

$$\max_{\{\beta_{0j},\beta_j\}_1^K \in \mathbb{R}^{K(p+1)}} \{\frac{1}{N}\sum_{i=1}^{N} \log p_{g_i}(x_i) - \lambda \sum_{j=1}^{K} P_\alpha(\beta_j)\}$$

where $p_j(x_i) = P(G = j|x_i)$, and $g_i \in (1, 2, \ldots, K)$ is the ith response. Therefore, MLR-based classifiers are able to be constructed by adopting this package.

The SVM classifier was built using the e1071 package (Chang and Lin, 2011). In this package, the multi-class problem is approached via the "one-against-one" approach (Knerr et al., 1990). Consequently, there are $K(K\text{-}1)/2$ classifiers that are needed to be constructed for K classes. In this study, the SVM-based classifier was required to construct three classifiers due to the presence of three classes. More precisely, the training data was used to form the ith and jth classes and was able to deal with the following two-class classification problem.

$$\max_{w^{ij},\, b^{ij},\, \xi^{ij}} \{\frac{1}{2}(w^{ij})^T w^{ij} + C\sum_t (\xi^{ij})_t\}$$

subject to

$$(w^{ij})^T \phi(x_t) b^{ij} \geq 1 - \xi^{ij}, \text{ if } x_t \text{ in the } i\text{th class},$$
$$(w^{ij})^T \phi(x_t) b^{ij} \leq -1 + \xi^{ij}, \text{ if } x_t \text{ in the } j\text{th class},$$
$$\xi^{ij} \geq 0.$$

Following this, a voting strategy is adopted, the class with the maximum number of votes is considered to be the most probable one.

The DT-based classifier was implemented using the caret package (Therneau and Atkinson, 2018) of R. Specifically, the package mainly provides classification and regression trees (CART). Furthermore, the randomForest package (Liaw and Wiener, 2002) of R was also employed in this study to construct a random forest-based classifier. The package mainly provides an R interface using a Fortran program developed by Breiman (2001). Ensemble learning and bagging are the two important concepts used when creating the random forests. Furthermore, a random forest is a classifier consisting of a collection of tree-structured classifiers (Breiman, 2001). Therefore, according the voting results, we should be able to obtain the prediction for a specific data-set. In addition, RF provided the functions, that allow the evaluation of the effect of features during model construction. The mean decrease in accuracy and mean decrease in node impurity are provided by randomForest package (Liaw and Wiener, 2002). Note that the impurity is defined as

$$I(p) = 1 - \sum_{i=1}^{J} p_i^2$$

where p_i is the probability of correct classification.

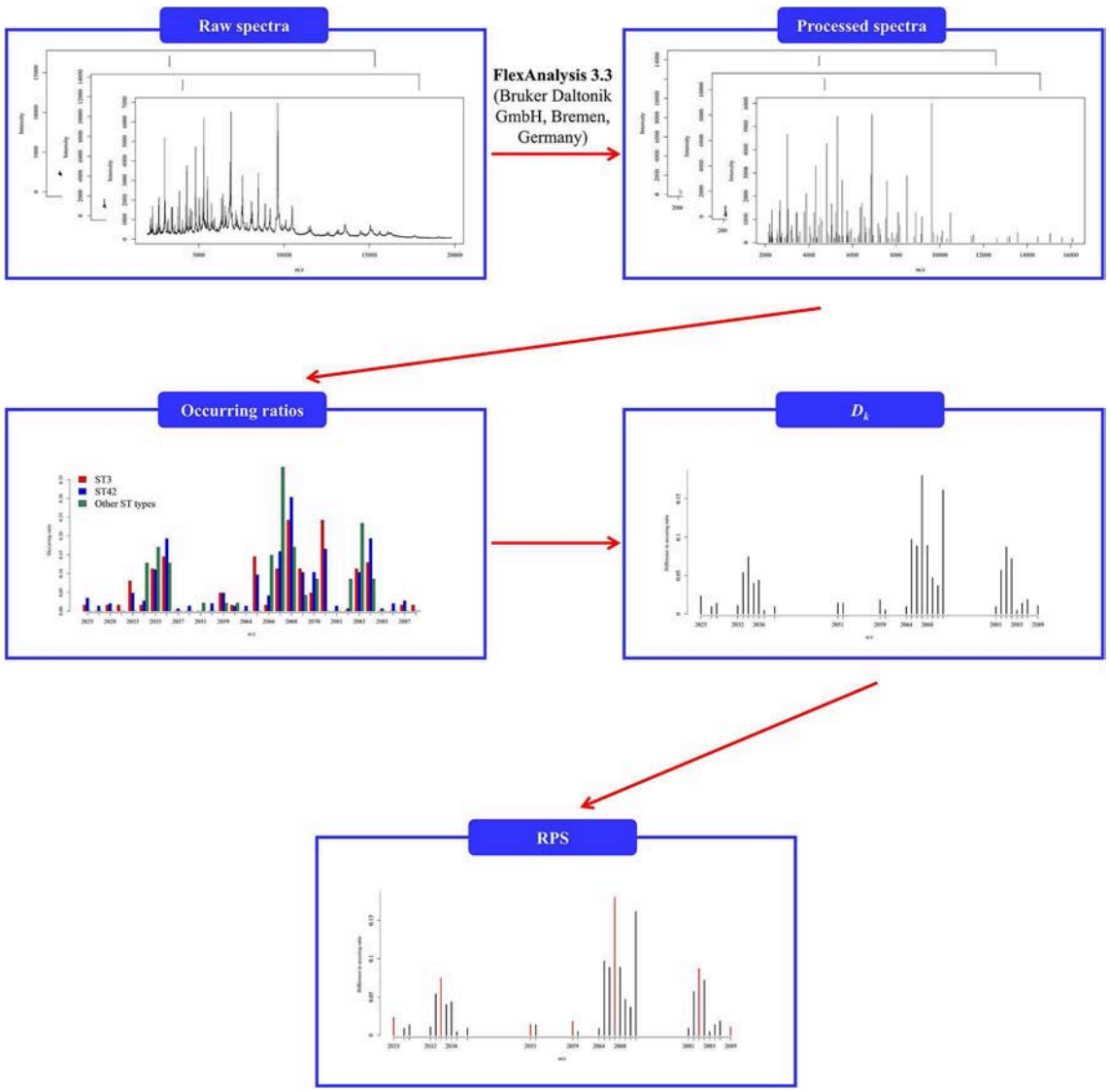


FIGURE 1 | Flowchart of preprocessing of spectral data given that the tolerance value is 5. The incidence ratio was determined by the number of the isolates among the CPS. D_k was defined as the total difference between the incidence ratios.

In addition to the aforementioned multiclass classification approaches, we also adopted these methods when examining binary classification in order to better distinguish ST3 and ST42. The same package was implemented for this process, but in this case using the binary option. For instance, logistic regression (LR) was used to construct the binary classification model using the glmnet package (Friedman et al., 2010). Similarly, for SVM, DT, and RF, the same packages were adopted.

Statistical Analysis

It is important to note that we were concerned not only with the frequency of the peaks, but also with the intensity of a specific peak among the multiple spectra, which is also a critical in discriminating these three groups. Therefore, in order to compare differences in intensities of specific peaks among these three groups, the Kruskal–Wallis test (Kruskal and Wallis, 1952) and Kendall's tau coefficient (Kendall, 1938) were both adopted as part of this study. Moreover, to obtain the ability of an individual peak to distinguish between the three groups, the area under the receiver operating characteristic curve (AUC) was taken into consideration. Note that to deal with multi-class performance evaluation, the pROC package (Robin et al., 2011) in R was implemented in order to obtain an estimation for the multi-class AUC (Hand and Till, 2001). When comparing the difference between two independent samples, the Wilcoxon rank-sum test was employed, and it was also implemented to compare cross validation performance. To find the optimal cut-off points for each ROC curve during binary classification,

the OptimalCutpoints package (López-Ratón et al., 2014) was applied.

Evaluation Metrics of the Classifiers

To evaluate the performance of the classifiers constructed by the aforementioned machine learning methods, the stratified 5-fold cross validation technique was implemented. The first procedure of the stratified 5-fold cross validation splits the dataset into 5 groups, preserving the percentage of data for each class. Then, one group is left as the testing dataset, while the remaining groups form the training dataset. The classification model was built according to the training dataset and was evaluated using the testing dataset. Note that each group was a testing dataset. Consequently, we obtained 5 prediction performances for these 5 groups. The average accuracy and the AUC among the five testing sets were determined in order to compare the performance when constructing the multiclass classifiers. As a result, the AUC was calculated by using the pROC package (Robin et al., 2011) in R. By way of contrast, we used sensitivity, specificity, accuracy, and AUC when evaluating the binary classification performance. More specifically, suppose that the class of ST42 is labeled as 1, these metrics are defined as follows:

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{FP + TN}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP means the true positives and refers to the number of ST42 that were correctly predicted by the classifier, TN means true negatives and refers to the number of ST3 that were correctly predicted by the classifier, FP means false positives and refers to the number of ST42 that were incorrectly predicted by the classifier, and FN means false negative and refers to the number of ST3 that were incorrectly predicted by the classifier.

Feature Selection Strategies

In addition to applying the importance evaluation from RF, we also developed two strategies, the stepwise strategy and the forward strategy, to find the peaks that needed to be considered as classifiers. More specifically, these two strategies were adopted when constructing the multi-class RF-based classifiers in order to obtain the peaks that are essential when distinguishing these three groups.

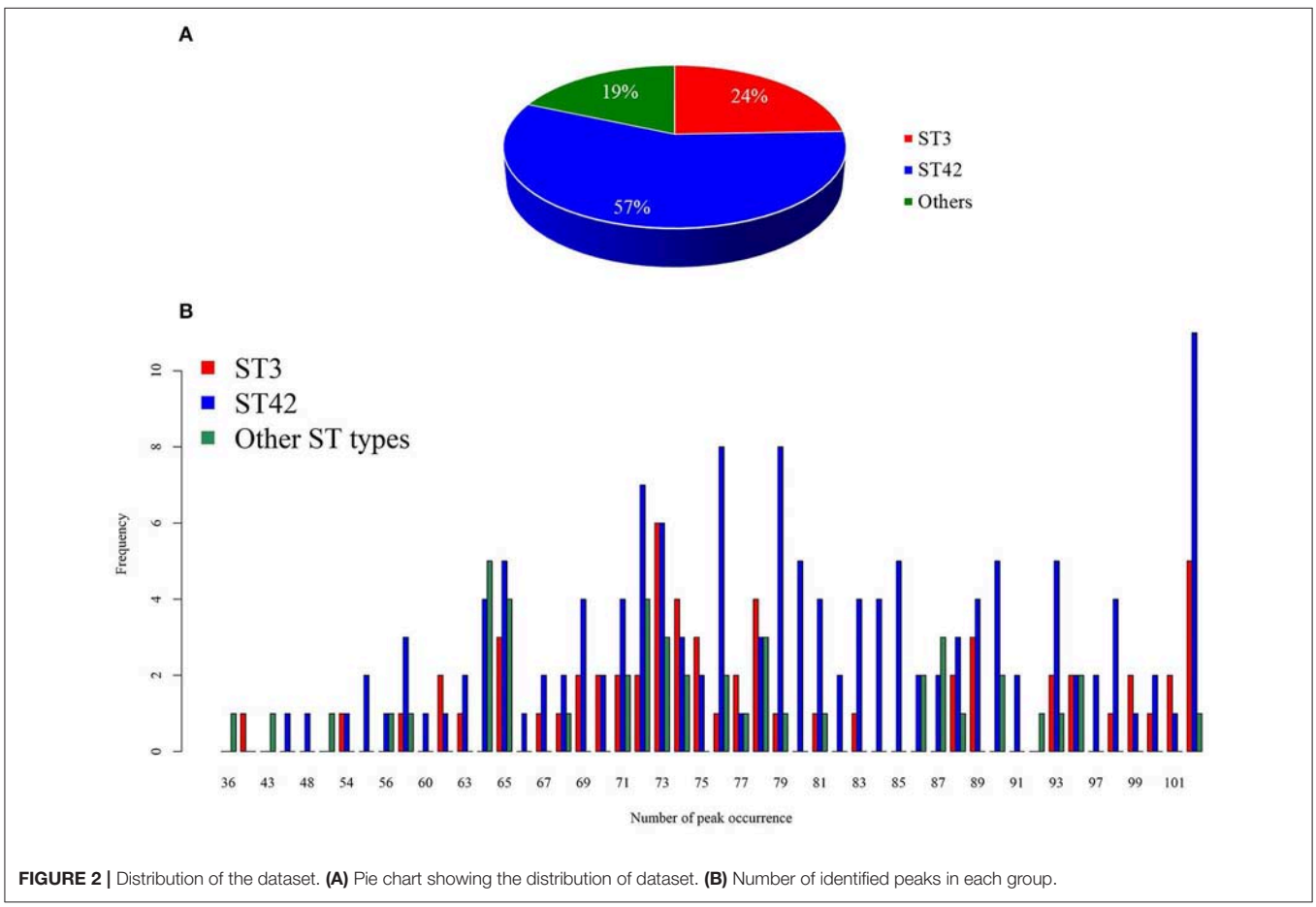


FIGURE 2 | Distribution of the dataset. **(A)** Pie chart showing the distribution of dataset. **(B)** Number of identified peaks in each group.

The stepwise strategy starts initially with a specific peak, such as the one with the largest AUC, the largest absolute value of Kendall's tau coefficient, and so on. Further, the next peak to be selected must attain the largest AUC or accuracy when combined with the currently selected peak(s) among those peaks that have not been selected. The process is then repeated until the AUC or the accuracy does not increase anymore.

When using the forward strategy, the peaks must be sorted into a specific order. For example, the peaks can be sorted by their AUCs in the descending order. Then the forward strategy would follow this order to adding new peaks if the new one is able to increase the AUC or accuracy. Otherwise, the peak will not be regarded as a helpful feature when constructing the classifier, and thus will be discarded.

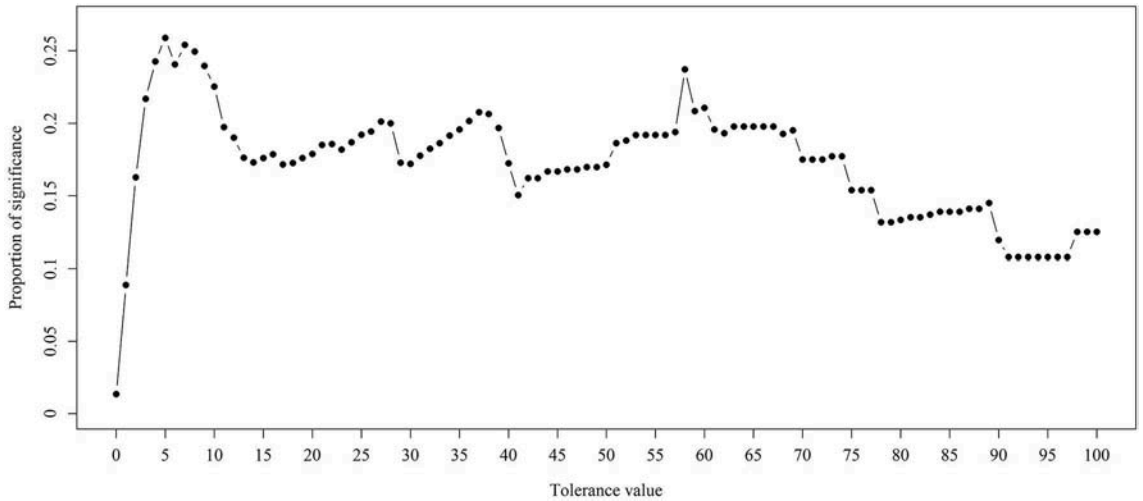


FIGURE 3 | Proportion of significance for different tolerance values. Fisher's exact test was employed to examine the difference between two different ST types. The *p*-values were derived by the average of three *p*-values.

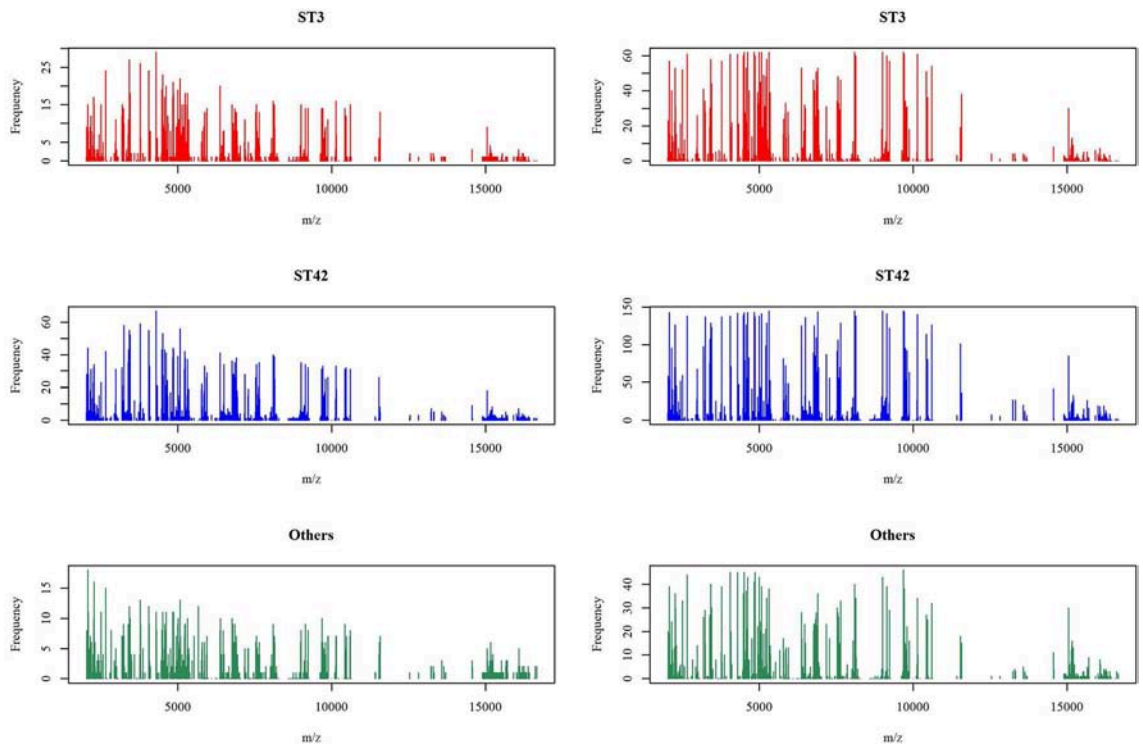


FIGURE 4 | Mass spectra before and after peak alignment. The left panel is the number of spectra appearing the specific peaks under the original signal of the mass spectra and the right panel is after the alignment strategy with tolerance value 5.

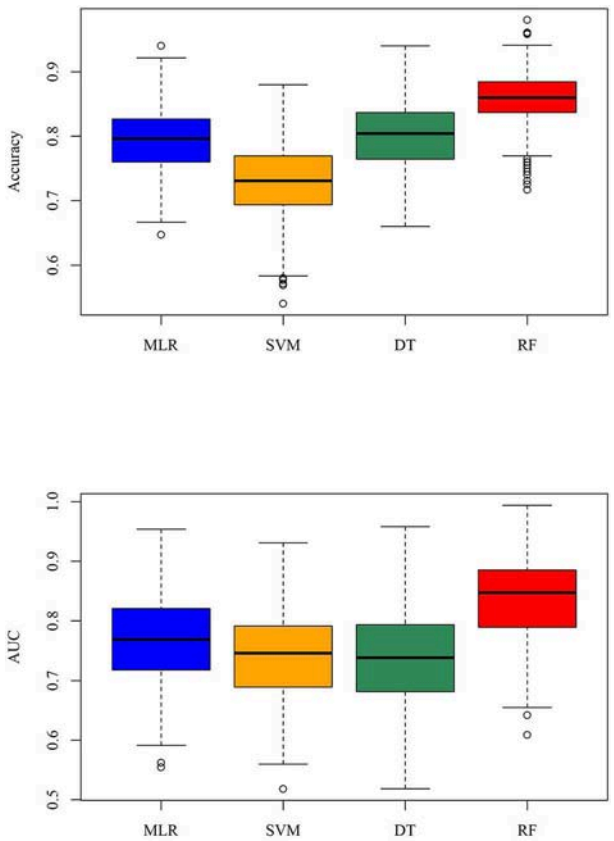


FIGURE 5 | Boxplot of the accuracy and AUC for the repeated 5-fold cross validation when the tolerance value is 5.

The sensitivity of both these strategies is dependent on the selection of the initial peak. In other words, the first selected peak will affect different peak combinations and this may produce different performances. Moreover, different criteria are likely also to result in different combinations. In this study, both AUC and the accuracy are two of the major concerns when building the multi-class classifiers. On the other hand, the balance between the sensitivity and specificity also needs to be taken into consideration. Nevertheless, the major aspects of the evaluation still are dependent on the AUC and the accuracy.

RESULTS

Summary Statistics of Spectra Data

Among the 254 isolates used in the present study, 62 isolates were ST3, 145 isolates were ST42, and 47 isolates were neither ST3 nor ST42 and formed a separate group of strains. The details of the other ST types show in **Supplementary Table 1**. Given that we aimed to develop and validate a rapid *S. haemolyticus* strain typing tool, we designed the classes based on the local epidemiology, whereas ST3 and ST42 accounted for the majority of strains. In clinical practice, the developed tool would provide preliminary strain typing information, notifying clinical physicians if the isolate of interest is of the major ST types. When the isolate of interest is classified by the model as a major ST type, outbreaks from the origin should be suspected and further investigation could be initiated immediately. As noted, this classification was determined by the local epidemiology of *S. haemolyticus* in Taiwan. **Figure 2** demonstrates the data statistics and the distribution of number of peaks identified for each group. On an average, the number of peaks identified in the range 2,000 Da to 17,000 Da was 76.48, with a standard deviation of 13.46. More specifically, the average number of peaks identified for ST3 was 77.03, while that of ST42 was 77.68, and the number of peaks identified for the other ST types was 72.04. Although the number of peaks identified for the other ST types seemed to be lower than that for the other two strains, the Kruskal-Wallis rank sum test did not show a significant difference between the three groups ($p = 0.0586$). When spectra signal intensity was examined, the average (standard deviation) normalized intensity across the three groups was 0.16 (0.18). The average normalized intensity of ST3 was 0.13 (0.16), while that of ST42 was 0.17 (0.19), and that of the other group of ST types was 0.18 (0.18). The normalized intensity of ST3 seemed to be lower than that of other two groups and the result of the Kruskal-Wallis rank sum test also showed that there were significant differences between these three groups ($p < 0.0001$).

Determination of Tolerance Value

In the previous section, we have described the strategy for determining the RPS using Fisher's exact test. **Figure 3** demonstrates the proportion of significance for different tolerance values. More specifically, the proportion of significance was determined by the number of occurring significance. Note that the significance here indicates that the *p*-value of Fisher's exact test is <0.0001. When the tolerance value is 5, the proportion of significance is highest. The spectra with and without preprocessing is shown in **Figure 4**. In addition, **Figure 5** demonstrates the performance of the 5-fold cross validation repeated 100 times. Specifically, there were 500 independent tests of ACCs and AUCs for evaluating whether the tolerance value was robust enough. These results implied that the tolerance value was adequate for further analysis. The AUC of different classifiers under different tolerance values, which are shown in **Figure 6**, demonstrated that the AUC was able to attain a value of 0.8 with a low standard deviation for the tolerance value of 5. Therefore, we used a tolerance value 5 for the feature selection because of its robustness. **Table 1** shows the mean ± standard deviation of the accuracy and AUC values for the 5-fold cross validation using the different machine learning methods. Wilcoxon rank sum test was then used to compare their performances. It should be noted that the *p*-value next to the accuracy/AUC column is from the Wilcoxon rank sum test results and this was employed to compare the accuracy/AUC when using the MLR method on the test data during 5-fold cross validation. Furthermore, we also found that the RF values tended to be robust due to the presence of a lower standard deviation compared to other methods for the different tolerance values present in **Figure 5**. Hence the feature selection strategies, when implemented to find important features, used RF. It should be noted that the number of peaks in RPS was 583 for a tolerance

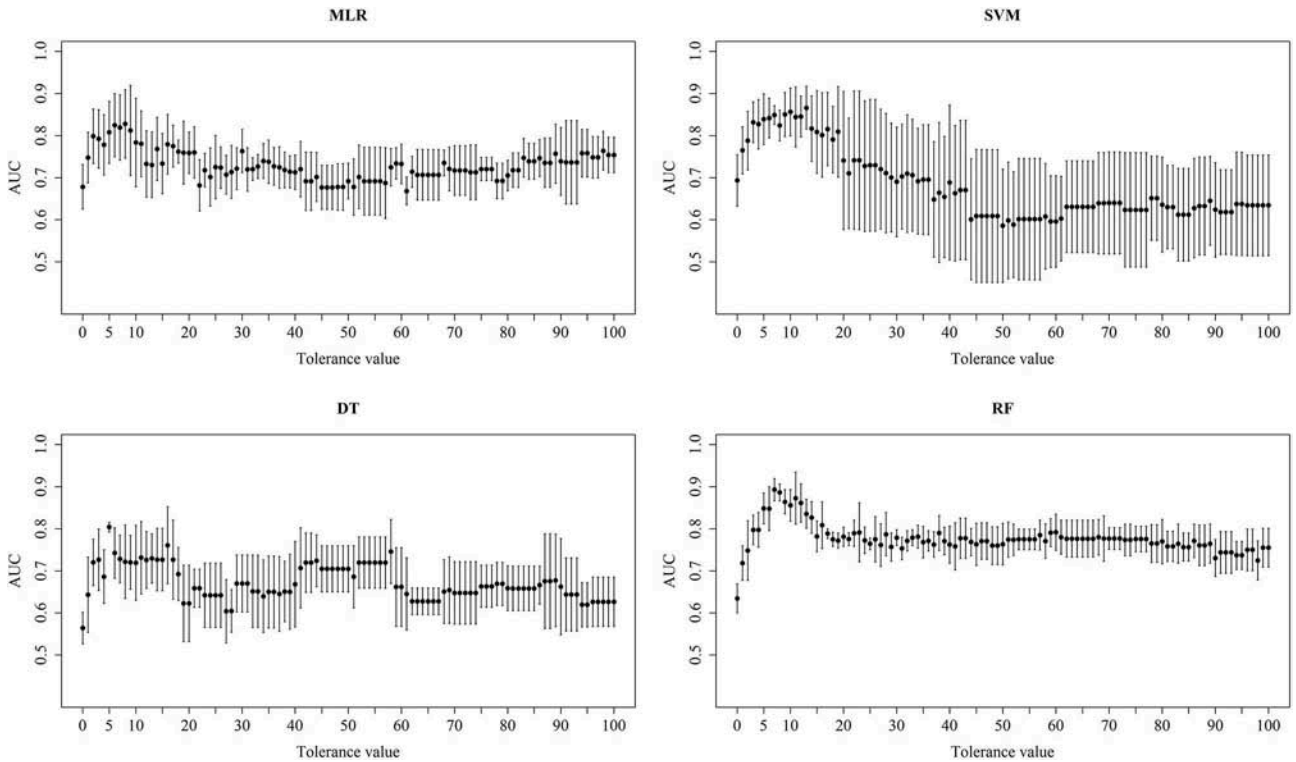


FIGURE 6 | Performance of different classifiers. Mean and standard deviation AUC of the 5-fold cross validations for the different tolerance values using different machine learning methods.

TABLE 1 | Performance of 5-fold cross validation.

	Accuracy	*p*-value	AUC	*p*-value
MLR	0.819 ± 0.028	–	0.808 ± 0.074	–
SVM	0.858 ± 0.029	0.0937	0.839 ± 0.060	0.5476
DT	0.840 ± 0.046	0.4005	0.804 ± 0.012	0.6905
RF	0.866 ± 0.014	0.0196	0.848 ± 0.037	0.3095

Mean ± standard deviation accuracy and AUC of the 5-fold cross validations for the multiclass classifications using different machine learning methods when the tolerance value is 5. The p-values were derived by comparing with MLR. MLR, multiclass logistic regression; SVM, support vector machine; DT, decision tree; RF, random forest.

value of 5 and thus it was these 583 features that were used to construct the multi-class classifiers used to discriminate the three groups.

Results of Feature Selection Strategies on RF-Based Classifiers

Table 2 demonstrates the results of the two feature strategies when RF was used to construct the classification models. The forward strategy was highly dependent on the order of inclusion of the features. On the other hand, the starting peak in the stepwise strategy was critical. Both these strategies demonstrated that a reduction in the number of features appeared to increase the accuracy or AUC. In other words, the selected peaks were found to be highly correlated with *S. haemolyticus* and were able to distinguish between the three groups of ST strains.

A total of 10 models were constructed by adopting different feature selection strategies and selecting different peaks. We next identified the peaks that were selected in more than five models and these were regarded as discriminative peaks. **Table 3** shows the occurrence and proportions of these discriminative peaks. From this table it can be seen that the ST42 isolates almost always present the peaks 4999 and 6496, explicitly they were present in over 90% of samples. However, neither ST3 nor ST42 ever presented the peak 5635. In addition, **Figure 7** presents the whole spectral incidence for the three groups, and specifically focuses on the area from 4700 to 7100 Da, which allows closer observation of the behavior of the discriminative peaks. Specifically, the red bars show the differences between these three groups that seem to be critical to constructing the classifiers. When considering the intensity, **Table 4** presents the means and standard deviations of the normalized intensities of the discriminative peaks. Since the incidence tends to be small, and the normalized intensity is between 0 and 1, the average values also tend to be low. Nevertheless, some peaks still showed strong intensity. For example, peaks 6781, 6496, and 4999 still have relatively large intensity values. The Kruskal-Wallis test was employed to test difference among the three groups and when there was a difference between two groups, the *p*-value tended to be lower. Hence the *p*-values in **Table 4** are very small. It should be noted that these discriminative peaks are the ones that

TABLE 2 | Performance of feature selection.

Strategy	Start up	# peaks	RF		*p*-value
			Accuracy	AUC	
Stepwise	AUC	21	0.918 ± 0.024	0.921 ± 0.025	0.0079
	Kendall's tau	26	0.906 ± 0.008	0.897 ± 0.049	0.2222
	KW	20	0.902 ± 0.030	0.917 ± 0.042	0.0952
	IMP	27	0.910 ± 0.008	0.926 ± 0.020	0.0079
Forward	AUC	18	0.897 ± 0.032	0.896 ± 0.026	0.0952
	Kendall's tau	35	0.901 ± 0.024	0.893 ± 0.047	0.2222
	FE	25	0.902 ± 0.023	0.898 ± 0.035	0.0556
	KW	26	0.906 ± 0.031	0.902 ± 0.061	0.2222
	IMP-ACC	22	0.882 ± 0.032	0.864 ± 0.059	0.7533
	IMP-GINI	28	0.874 ± 0.034	0.836 ± 0.037	0.5476
No		583	0.866 ± 0.014	0.848 ± 0.037	-

Mean ± standard deviation accuracy and AUC of the 5-fold cross validation of RF using different numbers of peaks selected by the forward and stepwise feature selection strategies using different orders of peaks and the corresponding performance using RF. AUC, area under the curve; FE, Fisher's exact test; KW, Kruskal-Wallis test; IMP-ACC, importance measure calculated by mean decreased accuracy using RF; IMP-GINI, importance measure calculated by mean decreased impurity using RF.

TABLE 3 | Number of occurrence peaks (proportions) and average *p*-values using the Fisher's exact test for the discriminative peaks.

Peak	Type 3	Type 42	Others	*p*-value
4673	40 (0.645)	5 (0.034)	5 (0.106)	0.022
5129	49 (0.790)	25 (0.172)	18 (0.383)	0.001*
4999	62 (1.000)	138 (0.952)	35 (0.745)	0.035
5635	0 (0.000)	0 (0.000)	12 (0.255)	0.333
6466	31 (0.500)	3 (0.021)	23 (0.489)	0.333
2499	52 (0.839)	59 (0.407)	33 (0.702)	0.035
3390	15 (0.242)	107 (0.738)	27 (0.574)	0.015
3411	20 (0.323)	70 (0.483)	1 (0.021)	0.015
5036	43 (0.694)	43 (0.297)	17 (0.362)	0.157
6496	30 (0.484)	136 (0.938)	15 (0.319)	0.039
6781	21 (0.339)	129 (0.890)	26 (0.553)	0.011

**Indicated that the $p < 0.01$.*

are often selected using the various different feature selection strategies shown in **Table 2**. Moreover, the boxplots in **Figure 8** can be used to demonstrate the distribution of intensities among the different ST types. According to **Table 3**, the intensity in event of a lower incidence tends to be smaller. This can also be seen in **Figure 8** for peaks such as 4674 and 4659.

Classifier for Discriminating ST3 and ST42

Table 5 shows the performance of the classifiers used to distinguish ST3 and ST42. Since the majority of data available was for ST42, the specificity of these classifiers tended to be higher. Even so, the AUCs among the different classifiers also showed impressive results. In both **Figures 7, 8**, it can be seen that the incidence and intensities are evidently different for some specific peaks.

DISCUSSION

This is a study that focused on the strain typing of *S. haemolyticus* based on the MALDI-TOF MS utilizing statistical tests and machine learning methods simultaneously. Specifically, the Fisher's exact test was employed to determine the reasonable tolerance values on preprocessing the spectra data. We have not only constructed machine learning-based classifiers that allow for different feature selection strategies, but have also employed statistical tests to compare the performance of the various discriminative peaks related to the different ST types. The rapid identification of *S. haemolyticus* strain types will facilitate the identification of origins of infection and will also provide critically-ill patients with substantial benefits because it will allow for rapid infection control. Additionally, further exploration of the discriminative peaks will allow the identification of each corresponding peptide. Such findings should provide clinically valuable information pertaining to the different subtypes of *S. haemolyticus*.

Previous studies used "type templates" for each ST type based on the incidence of specific peaks in their MALDI-TOF MS spectra in order to handle the issue of peak shifting; furthermore, log-transformed intensity was used to represent corresponding signal strength for each peak (Wang et al., 2018a,b). These studies also used the signals with the highest incidence probability in a local region (± 5 m/z) as the center of each peak feature. In other words, determining the local region was based on the incidence probability without the adoption of any statistical tests. In this study, we used statistical analysis and also measured the performance of classifiers. Such an approach involving measurement of the tolerance value is an excellent approach for dealing with the peak shift problem present when using spectral data. As the tolerance value increases, the number of peaks in the RPS decreases, and vice versa. The reason is that a larger tolerance value may lead to the alignment of more discriminative peaks with the same specific peak. In contrast, a lower tolerance value results in a paucity of data. Specifically, in these circumstances, much less data can be aligned to the same specific peak, which produces a reduced amount of training data and eventually results in poor performance. In such circumstances we used both Fisher's exact test, and an evaluation of the variation in performance of different classifiers with different tolerance values. In short, the variation among different classifiers and tolerance values was taken into consideration and this increased the robustness of our model. When the tolerance value was 5, the significance value was the largest and the standard deviation among the 5-fold cross validation analysis tended to be lower. Therefore, we used 5 as the final tolerance value when creating the RPS using 583 peaks.

There are a variety of machine learning methods that can be used for modeling different types of data. In this study, we adopted a number of relatively uncomplicated models to construct the classifiers. These uncomplicated methods are readily interpreted, which makes interpretation of the peak results easier and allows the initiation of further investigations into specific peaks simpler. Multinomial logistic regression is a generalized logistic regression model that is used for

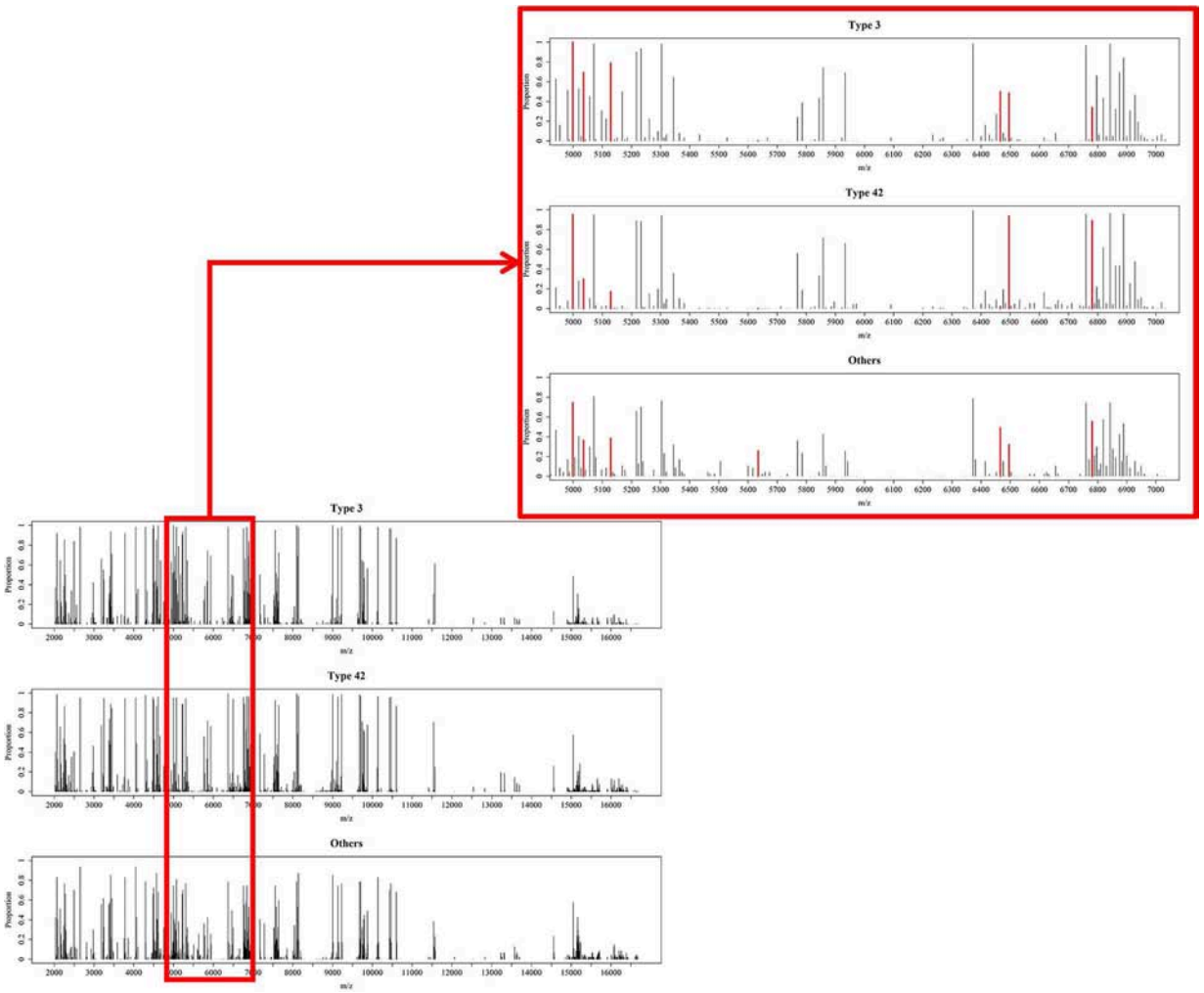


FIGURE 7 | Overview of processed MS data. Occurrence proportions among the three groups over the range from 2,000 to 17,000 Da and zoomed in for the range 4,900 to 7,100 Da. The red areas include peaks 4548, 4673, 4999, 5036, 5129, 5635, 6466, 6496, and 6781, which are the important peaks when constructing the RF-based classifiers.

handling multi-class problems and is one of the most common parametric statistical models. Our major concern in adopting the multinomial logistic regression model was multicollinearity. When the dependency among different independent variables is high, the estimators can be misinterpreted, and this may increase the prediction bias (Myers and Myers, 1990). Although the performance of MLR, as shown in **Table 1**, tended to be lower than other methods, the estimation of the parameters does seem to provide some information about the discriminative peaks. In other words, the estimators of the MLR were able to reveal which peaks potentially correlated with different ST types. It should be noted that a consideration of the standard errors of these estimators is an important reference point that can be used to avoid the multicollinearity effects. This is because there are few restrictions on the use of non-parametric methods such as SVM, DT, and RF. Their primary weakness is the time required for training the model when they use large scale datasets. However, this was not an issue in this study due to the relatively limited amount of data. Consequently, the performance of the non-parametric methods was better than that of MLR. Furthermore, the performance of RF was more robust than other methods. This is possibly due to two of the essential concepts of RF, namely ensemble learning and bagging. Previous studies also have reported the various advantages of RF (Boulesteix et al., 2012). In this study, we have also demonstrated that RF not only provides the highest accuracy and AUC, but it also retains the lower standard deviation.

Only a slight variation at the bacterial subspecies level is observed when they are compared using mass spectra (Lasch et al., 2014; Wang et al., 2018b). Nevertheless, until now, no studies have been able to identify the discriminative peaks when discriminating the different ST types of *S. haemolyticus* based on MALDI-TOF MS spectral data. Therefore, we used a variety of different strategies in order to identify the discriminative

peaks that are very likely to be highly related to the different ST types. An exploration of the discriminative peaks is highly dependent on the feature selection strategy and the machine learning method. It is important to note that the performance of RF is relatively robust and that it is also less time-consuming during training; in these circumstances, we largely adopted feature selection using RF for this study. The stepwise strategy is similar to the brute force method when used to find the best combinations for the classifiers. Consequently, the results of the stepwise strategy are generally better than those of the forward strategy. Furthermore, there is only one model that did not include peak 4673, which strongly supports peak 4673 as a discriminative peak. In addition, peak 5129 was not selected by two models, as shown in **Figure 8**, indicating that the normalized intensities across the three groups for this peak are apparently different. In addition, both **Figure 8** and **Table 3** show that the occurrence ratio is also significantly different across the three groups. Specifically, ST42 rarely presented peaks 4673 and 5129, while ST3 usually presented peaks at m/z 4673 and 5129. Further experiments are needed to identify the peptides corresponding to these peaks.

Although the machine learning-based classifiers has demonstrated impressive performance in this study for distinguishing different ST types of *S. haemolyticus*, there are

TABLE 4 | Means (standard deviation) and *p*-values using the Kruskal–Wallis test for the discriminative peaks.

Peak	Type 3	Type 42	Others	*p*-value
4673	0.052 (0.049)	0.003 (0.026)	0.011 (0.032)	<0.001*
5129	0.209 (0.200)	0.031 (0.094)	0.094 (0.150)	<0.001*
4999	0.769 (0.307)	0.350 (0.236)	0.455 (0.427)	<0.001*
5635	0.000 (0.000)	0.000 (0.000)	0.126 (0.266)	<0.001*
6466	0.145 (0.185)	0.003 (0.025)	0.118 (0.155)	<0.001*
2499	0.151 (0.120)	0.057 (0.083)	0.222 (0.207)	<0.001*
3390	0.030 (0.062)	0.132 (0.110)	0.103 (0.111)	<0.001*
3411	0.024 (0.040)	0.054 (0.065)	0.002 (0.011)	<0.001*
5036	0.115 (0.102)	0.032 (0.082)	0.076 (0.109)	<0.001*
6496	0.108 (0.151)	0.338 (0.210)	0.089 (0.174)	<0.001*
6781	0.065 (0.119)	0.247 (0.162)	0.108 (0.126)	<0.001*

**Indicated that the $p < 0.01$.*

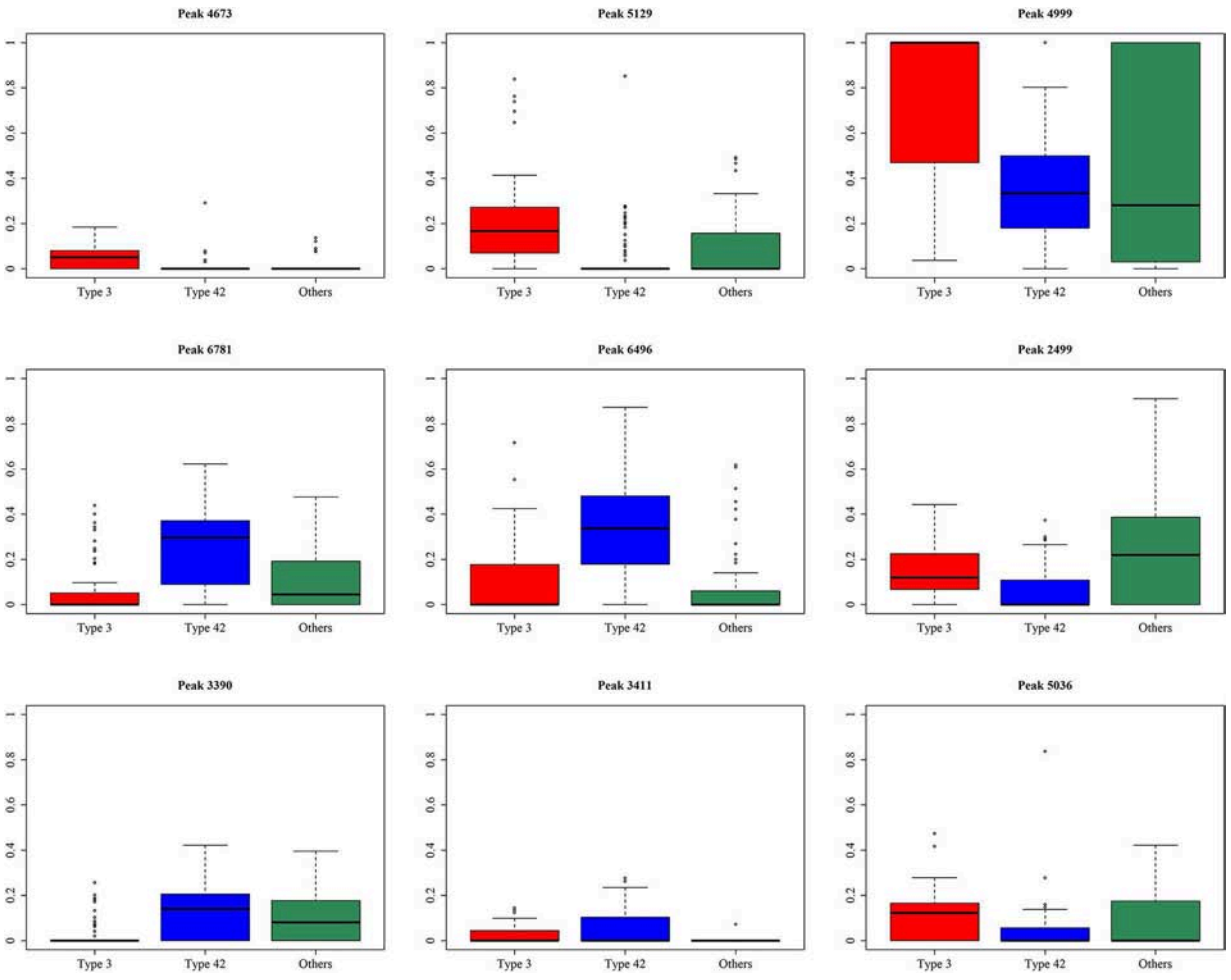


FIGURE 8 | Boxplots for the normalized intensity for the discriminative peaks.

TABLE 5 | Performance of binary classifier.

	Sensitivity	Specificity	Accuracy	AUC
LR	0.890 ± 0.062	0.968 ± 0.044	0.913 ± 0.045	0.919 ± 0.051
SVM	0.931 ± 0.055	0.983 ± 0.037	0.947 ± 0.032	0.969 ± 0.021
DT	0.938 ± 0.037	0.904 ± 0.032	0.928 ± 0.033	0.919 ± 0.024
RF	0.951 ± 0.031	1.000 ± 0.000	0.966 ± 0.022	0.972 ± 0.020

Mean ± standard deviation sensitivity, specificity, accuracy, and AUC of 5-fold cross validation for binary class classification using different machine learning methods when tolerance value is 5. LR, logistic regression; SVM, support vector machine; DT, decision tree; RF, random forest.

still some limitations. One major concern is that subspecies composition of the microbial strains may differ in different bacterial populations or in different regions of the world. In such circumstances the construction of machine learning-based classifier-based method might break down because these groups have different discriminative peaks for these subspecies. Even so, the machine learning-based classifier approach, in conjunction with the associated statistical tests, still provides a novel framework for analyzing MALDI-TOF MS data. Another critical issue that has been identified in the previous studies is the reproducibility of the mass spectra when MALDI-TOF MS is being used in bacterial typing (Walker et al., 2002; Wolters et al., 2011; Croxatto et al., 2012; Sandrin et al., 2013). There are a variety of factors involved in the reproducibility of the mass spectra and these include sample processing and specimen type (Josten et al., 2013; Sandrin et al., 2013; Mather et al., 2016). As of yet no standard protocol has been proposed for strain typing by MALDI-TOF MS. Nevertheless, a standard protocol should be optimized and specified for each species in order to achieve a robust performance when strain typing (Walker et al., 2002; Sandrin et al., 2013). The College of American Pathologists accreditation and proficiency test has been conducted for years to ensure the performance and quality standards of personnel and tests at Chang Gung Memorial Hospital, Linkou Branch. Therefore, on the basis of previous qualified MALDI-TOF MS workflow and data used here, the constructed classification models used in this study are readily available for *S. haemolyticus* strain typing.

Our study has demonstrated a method of developing robust classifiers for discriminating different ST types of *S. haemolyticus* based on MALDI-TOF MS data. The multi-class classifier demonstrated an AUC of 0.848 and accuracy of 0.886 when discriminating these three groups. If we only consider binary classification for ST3 and ST42, the AUC reaches an excellent discrimination power of 0.972. The constructed classifiers were able to provide instant information when identifying the origin of infection, which will allow rapid infection control. As a result, we believe that we have hereby developed a cost effective and rapid identification method for the strain typing of *S. haemolyticus*. This provides a great opportunity for further improvement of this new protocol and its introduction into routine clinical microbiology laboratory practices in order to attain rapid infection control. Furthermore, the explicit strategy for the determination of representative peaks before constructing the classifiers provides some indications for those who are interested in further analysis of spectra data.

DATA AVAILABILITY

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

H-YW carried out the data collection and curation. C-RC participated in the data analyses, model construction, and drafted the manuscript. C-RC, H-YW, FL, Y-JT, C-HC, T-PL, and T-YL participated in the design of the study and performed the draft revision. J-TH, T-YL, and J-JL conceived of the study, and participated in its design and coordination and helped to revise the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Ministry of Science and Technology, Taiwan (107-2320-B-182A-021-MY3 and 108-2221-E-008s-043-MY3) and Chang Gung Memorial Hospital (CMRPG3G1722).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.02120/full#supplementary-material

REFERENCES

Becker, K., Heilmann, C., and Peters, G. (2014). Coagulase-negative staphylococci. *Clin. Microbiol. Rev.* 27, 870–926. doi: 10.1128/CMR.00109-13

Boulesteix, A. L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisc. Rev.* 2, 493–507. doi: 10.1002/widm.1072

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Camoez, M., Sierra, J., Dominguez, M., Ferrer-Navarro, M., Vila, J., and Roca, I. (2016). Automated categorization of methicillin-resistant *Staphylococcus aureus* clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry. *Clin. Microbiol. Infect.* 22, 161.e1-161.e7. doi: 10.1016/j.cmi.2015.10.009

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transac. Intell. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Chang, P.-H., Liu, T.-P., Huang, P.-Y., Lin, S.-Y., Lin, J.-F., Yeh, C.-F., et al. (2018). Clinical features, outcomes, and molecular characteristics of an outbreak of *Staphylococcus haemolyticus* infection, among a mass-burn casualty patient group, in a tertiary center in northern Taiwan. *J. Microbiol. Immunol. Infect.* 51, 847–855. doi: 10.1016/j.jmii.2018.07.004

Croxatto, A., Prod'hom, G., and Greub, G. (2012). Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiol. Rev.* 36, 380–407. doi: 10.1111/j.1574-6976.2011.00298.x

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33:1. doi: 10.18637/jss.v033.i01

Froggatt, J. W., Johnston, J., Galetto, D. W., and Archer, G. (1989). Antimicrobial resistance in nosocomial isolates of *Staphylococcus haemolyticus*. *Antimicrob. Agents Chemotherap.* 33, 460–466. doi: 10.1128/AAC.33.4.460

Gibb, S., and Strimmer, K. (2012). MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* 28, 2270–2271. doi: 10.1093/bioinformatics/bts447

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410. doi: 10.1001/jama.2016.17216

Hand, D. J., and Till, R. J. (2001). A simple generalisation of the area under the ROC Curve for multiple class classification problems. *Mach. Learn.* 45, 171–186. doi: 10.1023/A:1010920819831

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., et al. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* 25:65. doi: 10.1038/s41591-018-0268-3

Hiramatsu, K. (1998). Vancomycin resistance in staphylococci. *Drug Resist. Updates* 1, 135–150. doi: 10.1016/S1368-7646(98)80029-0

Jolley, K. A., Bray, J. E., and Maiden, M. C. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST. org website and their applications. *Wellcome Open Res.* 3:124. doi: 10.12688/wellcomeopenres.14826.1

Josten, M., Reif, M., Szekat, C., Al-Sabti, N., Roemer, T., Sparbier, K., et al. (2013). Analysis of the matrix-assisted laser desorption ionization–time of flight mass spectrum of *Staphylococcus aureus* identifies mutations that allow differentiation of the main clonal lineages. *J. Clin. Microbiol.* 51, 1809–1817. doi: 10.1128/JCM.00518-13

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* 30, 81–93. doi: 10.1093/biomet/30.1-2.81

Knerr, S., Personnaz, L., and Dreyfus, G. (1990). "Single-layer learning revisited: a stepwise procedure for building and training a neural network," in *Neurocomputing. NATO ASI Series (Series F: Computer and Systems Sciences), vol 68,* eds Soulié F. F., and Hérault, J. (Berlin; Heidelberg: Springer), 41–50. doi: 10.1007/978-3-642-76153-9_5

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47, 583–621. doi: 10.1080/01621459.1952.10483441

Lasch, P., Fleige, C., Stämmler, M., Layer, F., Nübel, U., Witte, W., et al. (2014). Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of *Enterococcus faecium* and *Staphylococcus aureus* isolates. *J. Microbiol. Methods* 100, 58–69. doi: 10.1016/j.mimet.2014.02.015

Liaw, A., and Wiener, M. (2002). Classification and regression by random forest. *R News* 2, 18–22. Available online at: https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Lin, W.-Y., Chen, C.-H., Tseng, Y.-J., Tsai, Y.-T., Chang, C.-Y., Wang, H.-Y., et al. (2018). Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation. *Int. J. Med. Inform.* 111, 159–164. doi: 10.1016/j.ijmedinf.2018.01.002

López-Ratón, M., Rodríguez-Ñlvarez, M. X., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *J. Stat. Softw.* 61, 1–36. doi: 10.18637/jss.v061.i08

MacCannell, D. (2013). Bacterial strain typing. *Clin. Lab. Med.* 33, 629–650. doi: 10.1016/j.cll.2013.03.005

Mantini, D., Petrucci, F., Pieragostino, D., Del Boccio, P., Di Nicola, M., Di Ilio, C., et al. (2007). LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bioinformatics* 8:101. doi: 10.1186/1471-2105-8-101

Mather, C. A., Werth, B. J., Sivagnanam, S., Sengupta, D. J., and Butler-Wu, S. M. (2016). Rapid detection of vancomycin-intermediate *Staphylococcus aureus* by matrix-assisted laser desorption ionization–time of flight mass spectrometry. *J. Clin. Microbiol.* 54, 883–890. doi: 10.1128/JCM.02428-15

Myers, R. H., and Myers, R. H. (1990). *Classical and Modern Regression With Applications.* Belmont, CA: Duxbury Press.

Panda, S., Jena, S., Sharma, S., Dhawan, B., Nath, G., and Singh, D. V. (2016). Identification of novel sequence types among *Staphylococcus haemolyticus* isolated from variety of infections in India. *PLoS ONE* 11:e0166193. doi: 10.1371/journal.pone.0166193

Raymond, M., and Rousset, F. (1995). An exact test for population differentiation. *Evolution* 49, 1280–1283. doi: 10.1111/j.1558-5646.1995.tb04456.x

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Sandrin, T. R., Goldstein, J. E., and Schumaker, S. (2013). MALDI TOF MS profiling of bacteria at the strain level: a review. *Mass Spectrom. Rev.* 32, 188–217. doi: 10.1002/mas.21359

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., and Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: are we there yet? *Heart* 104, 1156–1164. doi: 10.1136/heartjnl-2017-311198

Therneau, T. M., and Atkinson, E. J. (2018). *An Introduction to Recursive Partitioning Using the RPART Routines.* Available online at: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

van Duin, D., Strassle, P. D., Dibiase, L. M., Lachiewicz, A. M., Rutala, W. A., Eitas, T., et al. (2016). Timeline of health care–associated infections and pathogens after burn injuries. *Am. J. Infect. Control* 44, 1511–1516. doi: 10.1016/j.ajic.2016.07.027

Vrioni, G., Tsiamis, C., Oikonomidis, G., Theodoridou, K., Kapsimali, V., and Tsakris, A. (2018). MALDI-TOF mass spectrometry technology for detecting biomarkers of antimicrobial resistance: current achievements and future perspectives. *Ann. Transl. Med.* 6:240. doi: 10.21037/atm.2018.06.28

Walker, J., Fox, A. J., Edwards-Jones, V., and Gordon, D. B. (2002). Intact cell mass spectrometry (ICMS) used to type methicillin-resistant *Staphylococcus aureus*: media effects and inter-laboratory reproducibility. *J. Microbiol. Methods* 48, 117–126. doi: 10.1016/S0167-7012(01)00316-5

Wang, H.-Y., Chen, C.-H., Lee, T.-Y., Horng, J.-T., Liu, T.-P., Tseng, Y.-J., et al. (2018a). Rapid detection of heterogeneous vancomycin-intermediate *Staphylococcus aureus* based on matrix-assisted laser desorption ionization time-of-flight: using a machine learning approach and unbiased validation. *Front. Microbiol.* 9:2393. doi: 10.3389/fmicb.2018.02393

Wang, H.-Y., Hsieh, C.-H., Wen, C.-N., Wen, Y.-H., Chen, C.-H., and Lu, J.-J. (2016). Cancers screening in an asymptomatic population by using multiple tumour markers. *PLoS ONE* 11:e0158285. doi: 10.1371/journal.pone.0158285

Wang, H.-Y., Lee, T.-Y., Tseng, Y.-J., Liu, T.-P., Huang, K.-Y., Chang, Y.-T., et al. (2018b). A new scheme for strain typing of methicillin-resistant *Staphylococcus aureus* on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using machine learning approach. *PLoS ONE* 13:e0194289. doi: 10.1371/journal.pone.0194289

Wang, H.-Y., Lien, F., Liu, T.-P., Chen, C.-H., Chen, C.-J., and Lu, J.-J. (2018c). Application of a MALDI-TOF analysis platform (ClinProTools) for rapid and preliminary report of MRSA sequence types in Taiwan. *PeerJ* 6:e5784. doi: 10.7717/peerj.5784

Wolters, M., Rohde, H., Maier, T., Belmar-Campos, C., Franke, G., Scherpe, S., et al. (2011). MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant *Staphylococcus aureus* lineages. *Int. J. Med. Microbiol.* 301, 64–68. doi: 10.1016/j.ijmm.2010.06.002

Wong, J. W. H., Cagney, G., and Cartwright, H. M. (2005). SpecAlign – processing and alignment of mass spectra datasets. *Bioinformatics* 21, 2088–2090. doi: 10.1093/bioinformatics/bti300

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., et al. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* 7:12474. doi: 10.1038/ncomms12474

Conflict of Interest Statement: 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.

Copyright © 2019 Chung, Wang, Lien, Tseng, Chen, Lee, Liu, Horng and Lu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 17 October 2019
doi: 10.3389/fmicb.2019.02346

Cross-Regional View of Functional and Taxonomic Microbiota Composition in Obesity and Post-obesity Treatment Shows Country Specific Microbial Contribution

Daniel A. Medina[1], Tianlu Li[2,3], Pamela Thomson[4], Alejandro Artacho[5], Vicente Pérez-Brocal[5] and Andrés Moya[5,6,7]*

[1] Laboratorio de Biotecnología Aplicada, Facultad de Medicina Veterinaria, Universidad San Sebastián, Puerto Montt, Chile, [2] Chromatin and Disease Group, Cancer Epigenetics and Biology Programme (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain, [3] Epigenetics and Immune Disease Group, Josep Carreras Leukaemia Research Institute (IJC), Barcelona, Spain, [4] Departamento de Ingeniería Química y Bioprocesos, Escuela de Ingeniería, Pontificia Universidad Católica de Chile, Santiago, Chile, [5] Genomics and Health Area, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO)-Salud Pública, Valencia, Spain, [6] Integrative Systems Biology Institute, University of Valencia, CSIC, Valencia, Spain, [7] Biomedical Research Centre Network for Epidemiology and Public Health (CIBEResp), Madrid, Spain

OPEN ACCESS

Edited by:
Qi Zhao,
Shenyang Aerospace University, China

Reviewed by:
Alexander V. Tyakht,
Institute of Gene Biology (RAS), Russia
Sofia Forslund,
Max Delbrück Center for Molecular Medicine, Germany

***Correspondence:**
Daniel A. Medina
daniel.medina@uss.cl

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

Received: *29 October 2018*
Accepted: *26 September 2019*
Published: *17 October 2019*

Citation:
Medina DA, Li T, Thomson P, Artacho A, Pérez-Brocal V and Moya A (2019) Cross-Regional View of Functional and Taxonomic Microbiota Composition in Obesity and Post-obesity Treatment Shows Country Specific Microbial Contribution. Front. Microbiol. 10:2346. doi: 10.3389/fmicb.2019.02346

Gut microbiota has been shown to have an important influence on host health. The microbial composition of the human gut microbiota is modulated by diet and other lifestyle habits and it has been reported that microbial diversity is altered in obese people. Obesity is a worldwide health problem that negatively impacts the quality of life. Currently, the widespread treatment for obesity is bariatric surgery. Interestingly, gut microbiota has been shown to be a relevant factor in effective weight loss after bariatric surgery. Since that the human gut microbiota of normal subjects differs between geographic regions, it is possible that rearrangements of the gut microbiota in dysbiosis context are also region-specific. To better understand how gut microbiota contribute to obesity, this study compared the composition of the human gut microbiota of obese and lean people from six different regions and showed that the microbiota compositions in the context of obesity were specific to each studied geographic location. Furthermore, we analyzed the functional patterns using shotgun DNA metagenomic sequencing and compared the results with other obesity-related metagenomic studies, we observed that microbial contribution to functional pathways were country-specific. Nevertheless, our study showed that although microbial composition of obese patients was country-specific, the overall metabolic functions appeared to be the same between countries, indicating that different microbiota components contribute to similar metabolic outcomes to yield functional redundancy. Furthermore, we studied the microbiota functional changes of obese

patients after bariatric surgery, by shotgun metagenomics sequencing and observed that changes in functional pathways were specific to the type of obesity treatment. In all, our study provides new insights into the differences and similarities of obese gut microbiota in relation to geographic location and obesity treatments.

Keywords: human gut microbiota, bariatric surgery, obesity, functional redundancy, metagenomic, functional convergence

INTRODUCTION

It has been established that the composition of human gut microbiota greatly influences human health (Carding et al., 2015, reviewed by Hall et al., 2017). Although the microbiota composition of each individual is relatively stable across adult life, it varies widely between individuals (Lozupone et al., 2012; Kolde et al., 2018). It has been observed that the human gut microbiota is primarily dominated by the phyla Firmicutes and Bacteroidetes (Zhernakova et al., 2016), with lesser contribution from Actinobacteria, Proteobacteria and Verrucomicrobia (Qin et al., 2010). Nevertheless, the microbiota composition differs geographically, primarily based on a variety of factors including host genetics, dietary habits, age, geographic location and lifestyle (Yatsunenko et al., 2012; Chilton, 2014; Nishijima et al., 2016; Fujio-Vejar et al., 2017; Gupta et al., 2017). To date, most of the studies on human gut microbiota have focused on populations from North America and Europe and although several of studies have demonstrated associations between microbiota alterations and diseases, including obesity, the specific contribution of these alterations to treatment response and how they differ across geographic locations still need to be envisioned.

Worldwide, obesity has nearly tripled since 1980 (Stevens et al., 2012). Information published by the World Health Organization (2016) shows that more than 1.9 billion adults, 18 years and older, have a body mass index (BMI) above 25 kg/m^2 among which, over 650 million have a BMI $>$ 30 kg/m^2, hence classifying them as obese. Overweight and obesity are defined as abnormal or excessive fat accumulation due to environmental and genetic factors (Angelakis et al., 2012). While overweight individuals haves a BMI range between 25 and less than 30 kg/m^2, obesity is classified by BMI as: obesity grade 1 (30 to 35 kg/m^2), grade 2 (35 to 40 kg/m^2) and grade 3 (40 to 60 kg/m^2) (NHLBI Expert Panel, 1998). Diverse studies have associated obesity with altered gut microbiota and reduced functional potential (Ley et al., 2006; Turnbaugh et al., 2006; Tremaroli and Backhed, 2012; Karlsson et al., 2013; Le Chatelier et al., 2013; Damms-Machado et al., 2015; Haro et al., 2016; Medina et al., 2017), which can be partially reversed after surgical intervention (Palleja et al., 2016). It has been originally observed that the relative abundance of Firmicutes and Bacteroidetes can be altered in obese patients, where an over-representation of Firmicutes is observed, in contrast to their lean counterparts (Ley et al., 2006; Turnbaugh et al., 2009; Million et al., 2012; Walters et al., 2014; Kasai et al., 2015; Haro et al., 2016). These taxonomic differences between lean and obese subjects may contribute to the development and perpetuation of obesity in several ways, including fat storage, regulation of energy metabolism, energy extraction from short chain fatty acids, increased low-grade inflammation and altered bile acid metabolism (Qin et al., 2012; Karlsson et al., 2013; Khan et al., 2016). However, a recent study focusing on the meta-analysis of previously published data using forest machine learning models showed no significant differences in gut microbiota composition between obese and healthy individuals (Sze and Schloss, 2016). Nevertheless, further research is required to unravel the exact role of gut microbiota composition in obesity context, and the factors, including geographical location, that may have an influence on not only differential microbial abundance but also long-term patient outcome.

Gastric surgical procedures, commonly known as bariatric surgery, have been successful in mediating long-term weight loss and reducing the incidence of related comorbidities (Sjöström, 2008; Eldar et al., 2011). Roux-en-Y gastric bypass (RYGB) is one of the most common bariatric surgery procedures in the United States (Smoot et al., 2006), where a small stomach pouch is connected to the proximal jejunum to directly bypass food to the small intestine, resulting in restrictive and malabsorptive nutrient intake (Tice et al., 2008; Tran et al., 2016). Another common bariatric surgery is sleeve gastrectomy (SG), where a significant portion of the stomach is removed to decrease its volume, leading to a significant reduction in the amount of food consumed (Gumbs et al., 2007). Since their invention, a number of studies have observed changes in obesity-associated microbiota and functional gene richness following different types of bariatric surgery (Zhang et al., 2009; Graessler et al., 2013; Kong et al., 2013; Tremaroli et al., 2015; Shao et al., 2016; Ilhan et al., 2017; Medina et al., 2017; Murphy et al., 2017). Specific changes following surgical intervention include an increase in Proteobacteria (*Escherichia coli*, *Enterobacter* spp.), changes in *Bacteroides* and *Prevotella* abundance, accompanied by an increase in *Akkermansia* and a decline in *Clostridium* genus, and global changes in Firmicutes/Bacteroidetes phyla ratios (Zhang et al., 2009; Furet et al., 2010; Li et al., 2011; Kong et al., 2013; Damms-Machado et al., 2015; Louis et al., 2016; Palleja et al., 2016; Ilhan et al., 2017; Medina et al., 2017). These microbial changes following bariatric surgery have shown to mediate different outcomes and, interestingly, such changes appear to depend on the type of surgery performed (Ilhan et al., 2017; Medina et al., 2017; Murphy et al., 2017). Specifically, metagenomic studies following RYGB showed an increase in pathways involving aerobic respiration and glutathione transfer and metabolism (Palleja et al., 2016). Both observations were in agreement with the increase in Proteobacteria phylum abundance, driven by facultative anaerobes, such as *E. coli*, as a result of lower gastric acid exposure during stomach transit. In the same study, the authors found an increase in

the pathways that degrade putrescine to succinate to produce gamma-aminobutyric acid (GABA) as a byproduct. GABA is known to act on receptors in the hypothalamus to promote satiety and, additionally, is thought to promote GLP-1 release, a positive regulator of GABA production by pancreatic beta-cells, which provides further involvement of this pathway following RYGB (Palleja et al., 2016). In addition, a complementary study conducted 9 years after RYGB intervention showed that bacterial rearrangements were stable in the long run and surgically altered microbiota promoted reduced fat deposition in recipient mice (Tremaroli et al., 2015).

In this study, we compared the gut microbiota composition and functional patterns of obese subjects from Chile with published data from Italy, Denmark, United States, France, and Saudi Arabia, in order to better understand the microbiota contribution to obesity. We found obese subjects to show geographic specificity in regards to relative microbiota abundance, similar to what has been previously observed in healthy individuals. Interestingly, the gut microbiota of obese patients did not display differential enrichment of functional pathways between countries, indicating that the geography-specific microbial compositions converge to perform similar functions in obese individuals. Furthermore, this study analyzed the metagenomic profiles of Chilean patients subjected to gastric bypass and sleeve gastrectomy and observed changes in the functional capacity of gut microbiota after surgery. Specifically, we identified *Akkermansia muciniphila* as one of the bacteria that drives the change in metabolic pathways after surgical intervention for obesity in Chilean patients.

MATERIALS AND METHODS

Sample and DNA Raw Data Collection

DNA raw data sequences from the stool of obese and lean subjects were obtained from studies carried out in Chile (Medina et al., 2017; Thomson et al., 2019), Italy (Tremaroli et al., 2015), Denmark (Palleja et al., 2016), United States (Ilhan et al., 2017), France and Saudi Arabia (Yasir et al., 2015), described in **Table 1** and **Supplementary Table S1**. In addition, 12 DNA stool samples from Chilean obese patients before and after treatment obtained from a previous study (Medina et al., 2017) were sequenced using shotgun metagenomics. All experiments were conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Faculty of Medicine, Pontifical Catholic University of Chile (n° 15-337).

Taxonomic Profiling From 16S rDNA Gene Amplicon Sequencing

Raw data of DNA sequences belonging to different studies were downloaded from ENA-EMBL or SRA-NCBI databases (**Supplementary Table S1**). To access the microbiota taxa and abundance, in this study we re-analyzed all raw data sequences using Microbiome Helper v2.3 OVA pipeline (Comeau et al., 2017). Briefly, MiSeq paired-end sequences were joined using PEAR (Zhang et al., 2014). Joined sequences were filtered by quality and length, in which demultiplexing and barcode depletion were performed using Microbiome Helper and QIIME v1.9.1 scripts (Caporaso et al., 2010; Navas-Molina et al., 2013; Comeau et al., 2017). Chimera sequences were filtered using the Vsearch tool (Rognes et al., 2016). Operational taxonomic units (OTUs) were picked by open-reference command and defined by clustering at 3% divergence (97% similarity) using the GreenGenes database release 08-2013 as reference (DeSantis et al., 2006; McDonald et al., 2012). Diversity analyses were performed using QIIME v1.9.1 scripts under Microbiome Helper v2.3 environment. The sequencing depth for even sub-sampling and maximum rarefaction depth was at least 10000 counts/sample, regard the minimum value obtained after OTU-picking for each data set (**Supplementary Table S2**). Predictive metagenomic functional profiling of microbial composition were performed using PICRUSt (Langille et al., 2013), following the instructions provided in their Metagenomic Prediction Tutorial. KEGG Orthology analyses were performed using the "ko_to_pathway_map" PICRUSt database to identify the enrichment of metabolic pathways.

Taxonomic abundance and metagenomic prediction tables were exported to R environment (R Core Team, 2013) for statistical analysis and figures were represented using the package LSD Lots of Superior Depictions (Schwalb et al., 2011). Volcano plots were constructed using the adjusted FDR *p*-value obtained from the unpaired *t*-test comparisons and the fold change of each condition. The taxonomic processed data obtained from QIIME and the metadata for the raw data files used in this study are exhibited in **Supplementary Table S2**.

The PICRUSt functional prediction was validated comparing the KEGG Orthology abundances obtained with the Uniref90 metagenomic abundances from the HUMAnN2 output. For this, the metagenomic Uniref90 table was converted to KEGG Orthology using the script humann2_regroup_table from HUMAnN2 (Abubucker et al., 2012). Both KEGG Orthology datasets were merged by row and normalized using quantile normalization to compare linear dependence between each data set using Pearson and pairwise Spearman rank correlation in the R environment (R Core Team, 2013).

Functional Annotation and Metagenomic Profiling of Fecal DNA

A total of 12 DNA stool samples were sequenced using the Illumina HiSeq next-generation sequencing platform carried out at Genoma Mayor (Universidad Mayor, Chile) with an output of 2×100 pb and 20×10^6 paired-end reads per sample. The raw data produced from DNA sequencing in this study were stored at the ENA-EMBL database under the accession number PRJEB29060. To evaluate the taxa abundance and metagenomic composition, all metagenomic raw sequences were analyzed in parallel using the Microbiome Helper v2.3 environment following the metagenomics SOP v2 tutorial. This pipeline allowed the calculation of microbiota abundance using functions from MetaPhlAn2 (Truong et al., 2015) and functional profile using HUMAnN2 (Abubucker et al., 2012). DNA sequence quality control was performed using KneadData (McIver et al., 2018), in which low quality sequences were first removed by Trimmomatic (Bolger et al., 2014), followed

TABLE 1 | Description of studies used in this work.

Study	Country	Methodology	N° lean subjects	N° obese subjects	DOI
Medina et al., 2017	Chile	16S rDNA Amplicon	0	19	doi: 10.7717/peerj.3443
Thomson et al., 2019	Chile	16S rDNA Amplicon	28	0	doi: 10.1017/S0007114519001570
Yasir et al., 2015	France	16S rDNA Amplicon	12	16	doi: 10.1038/nutd.2015.3
Yasir et al., 2015	Saudi Arabia	16S rDNA Amplicon	9	9	doi: 10.1038/ismej.2017.71
Ilhan et al., 2017	United States	16S rDNA Amplicon	10	15	doi: 10.1038/ismej.2017.71
Medina, this study	Chile	Shotgun metagenomics	0	6	This study
Palleja et al., 2016	Denmark	Shotgun metagenomics	0	13	doi: 10.1186/s13073-016-0312-1
Tremaroli et al., 2015	Italy	Shotgun metagenomics	0	7	doi: 10.1016/j.cmet.2015.07.009

by the application of Bowtie2 to screen out the contaminant DNA sequences mainly from human and viruses (Langmead and Salzberg, 2012). Both programs were run simultaneously for all samples using the GNU Parallel tool to repeat concatenate the entire process for each data set (First and Job, 2011). The paired-end files were merged using the script provided by the Microbiome Helper pipeline and then taxonomic and functional profiling were performed using MetaPhlAn2 and HUMAnN2, respectively. Gene family and pathway abundances were normalized for each sample and represented as a percentage. The resulting tables were exported to the R environment (R Core Team, 2013) for statistical analysis and figure representation. Metadata sets used for DNA shotgun sequencing to obtain stratified and unstratified metagenomic profiles are listed in **Supplementary Table S3**.

Data Comparison and Statistical Analysis

Statistical analyses were conducted using the R environment (R Core Team, 2013) version 3.4.4 (2018-03-15). Before calculations between countries, data sets were scaled proportionally, and using the R package preprocessCore (Bolstad, 2019), quantile normalization was performed to reduce batch effects between data belonging to different sources as was previously used for genome-wide analyses (Sun et al., 2011; Guo et al., 2014; Fei et al., 2018). For differential analysis of group variance between lean and obese microbiota abundance, we used Canonical Correspondence Analysis (CCA, a.k.a. constrained correspondence analysis) and Adonis tests, in which both were calculated using the R package vegan (Oksanen et al., 2018). The Adonis test was used for permutational multivariate analysis of variance using distance matrices, in which samples were fitted to linear models to calculate the whole compositional variability taking into account different sources of variation as well as the interactions between them. Spearman's rank correlation analyses were used to assess taxa abundance or functional abundance associations between the samples groups. To analyze differences in group means between lean and obese microbiota abundance, we applied Wilcoxon sum rank test with FDR adjustment utilizing the Benjamini Hochberg (BH) method. This was performed using the R command p.adjust(), in which we considered tests with FDR < 0.05 to be significant. In order to identify differentially enriched biomarkers among the compared groups, we applied the LEfSe analytic method using the online interface Galaxy[1] (Segata et al., 2011). Statistical differences between the KEGG Orthology abundances were calculated using the unpaired *t*-test and adjusted by BH FDR as described above. $-\log_{10}$ FDR was plotted against $\log_2$ ratio of each country in respect to Chile and represented using the R package LSD Lots of Superior Depictions (Schwalb et al., 2011).

RESULTS

Gut Microbial Diversity Is Specific to Geographic Locations

The composition and function of the human gut microbiota represent one of most important factors involved in obesity and its treatment (Ley et al., 2006; Tremaroli and Backhed, 2012; Karlsson et al., 2013; Le Chatelier et al., 2013; Damms-Machado et al., 2015; Medina et al., 2017). It has been previously established that the human gut microbiota displays pronounced differences between individuals residing in distinct geographic locations (Yatsunenko et al., 2012; Fujio-Vejar et al., 2017; Gupta et al., 2017). It is therefore of great interest to compare the microbial diversity of obese individuals from different geographical locations around the world. In this line of research, this study compared the gut microbial diversity of obese and lean subjects from Chile (Medina et al., 2017; Thomson et al., 2019) with data published by other studies in different regions around the world, namely United States (Ilhan et al., 2017), France and Saudi Arabia (Yasir et al., 2015), described in **Table 1**. Taxonomic microbiota abundance were collected by sequencing the 16S rDNA hypervariable regions V3–V4 or V4–V5 using the Illumina MiSeq platform (**Supplementary Table S1**), and raw data was processed as described in materials and methods. Taxonomic abundance comparisons at the genus level from the data obtained showed significant diversity in microbiota composition in lean people as previously reported (Yatsunenko et al., 2012; Nishijima et al., 2016; Fujio-Vejar et al., 2017; Gupta et al., 2017), but also is country specific in gut microbiota of obese subjects. Specifically, data variance analysis using CCA and Adonis tests showed significant differences (p-value < 0.05) in the gut microbiota composition between Chile, United States, France, and Saudi Arabia, in which the lean and obese microbiota were clustered

[1] http://huttenhower.sph.harvard.edu

by country (**Figures 1A,B**). Utilizing a different approach, pairwise Spearman rank correlation, we further demonstrated that the taxonomic abundance correlation of obese gut microbiota differed between countries (**Supplementary Figure S1A**). Nevertheless, although obese microbiota composition is clustered by country, there is significant inter-individual variability between subjects, indicated by high dispersion patterns observed in data variance analyses (**Figure 1**). In this regard, although some proximity is observed between France and Saudi Arabia microbiota, CCA and Adonis test variance analyses between the two countries shows a significant statistical difference (**Supplementary Figure S1B**). In all, these comparisons indicate that the microbiota compositions of obese patients and lean subjects were specific to each analyzed geographic location and displayed significant differences between countries. Comparing taxonomic abundance of obese individuals with their lean counterparts from the same country we observed significant differences by multivariate variance analyses for Chile, France, and Saudi Arabia (CCA p-value < 0.05); however, no statistically significant differences were observed for United States (**Supplementary Figure S2**). Wilcoxon sum-rank tests showed significant differences (FDR < 0.05) between lean and obese microbiota in Chile and France data sets, but not in United States and Saudi Arabia. Interestingly, although CCA showed clustering by country in both cases, France and Chile shared some genus changes between lean and obese microbiota (*Megasphaera*, *Veillonella*, *Adlercreutzia* and *Lachnospira*) (**Supplementary Table S2**). In addition, heatmap clustering by complete linkage method and the stacked bar plot of taxonomic abundances showed differential patterns not only between obese and lean subjects, but also between countries, which again highlights our observation that taxonomic microbial distribution is specific to each country (**Figure 2**).

Next, using the PICRUSt tool (Langille et al., 2013) designed to infer the functional patterns from taxonomic profile, we compared predicted KEGG Orthology (KO) enrichment of obese subjects from United States, France, and Saudi Arabia with the KO obtained from the obese Chilean data as reference in order to identify potential functional differences. Here, we observed significant differences (FDR < 0.05) in enriched pathways for all the countries studied with respect to Chile (**Figure 3**). Downregulated pathways in Saudi Arabia displayed an enrichment in translation and energy metabolism, whereas the United States and France showed a downregulation in membrane transport. For upregulated pathways, United States showed a specific enrichment in cellular processing and signaling and amino acid metabolism, whereas France and Saudi Arabia were found to enrich in pathways involving membrane transport (**Supplementary Figure S3**). In summary, the differences observed by functional prediction suggested that the human gut metagenome of obese subjects, like their healthy counterparts (Yatsunenko et al., 2012), may differ according to the geographical location.

Metagenomic Variations Across Regional Location in the Context of Obesity

Our initial analysis utilizing 16S rDNA amplicon sequencing showed marked differences between the compositions of the gut microbiota of obese subjects from different countries, therefore, to further validate our initial observation, we analyzed stool samples of 6 Chilean pre-treatment obese patients obtained from a previous study (Medina et al., 2017) by shotgun DNA

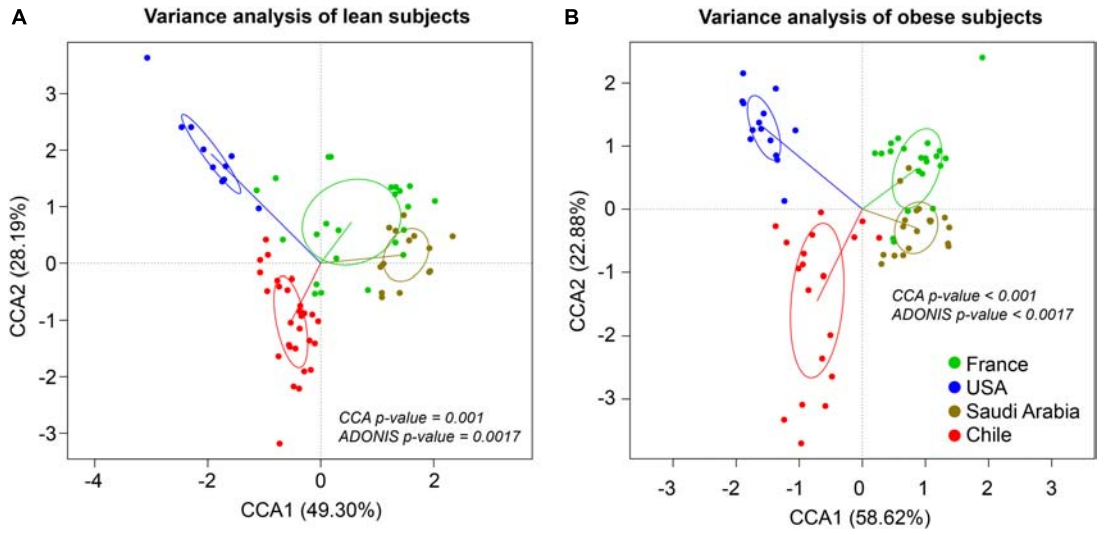


FIGURE 1 | Taxonomic abundance comparison at genus level from 16S rDNA sequencing. Constrained Correspondence Analysis (CCA) and Adonis test were performed to assess the variance in microbiota profiles at the genus level in lean **(A)** and obese **(B)** subjects. Vectors represent quantitative explanatory variables with confidence circles depicted for each country. Corresponding p-values are shown for each analysis and p-value < 0.05 was considered statistically significant. x- and y-axis show CCA1 and CCA2 components, respectively. Green, blue, brown and red points indicate individuals from France, United States, Saudi Arabia, and Chile, respectively.

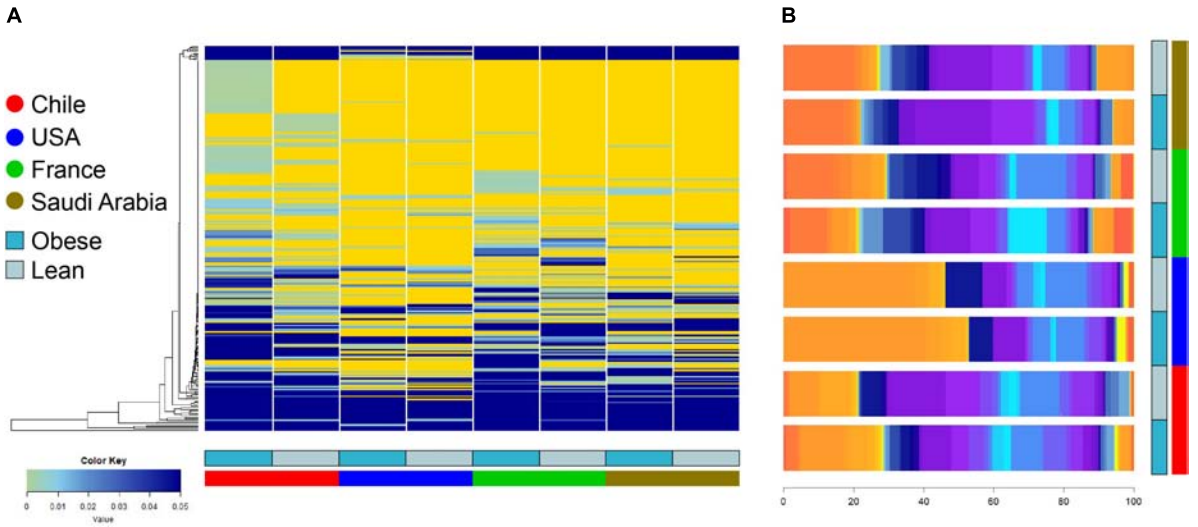


FIGURE 2 | Heatmap clustering and stacked bar plot of taxonomic abundances of lean and obese subjects at genus level. **(A)** Taxonomy abundance at genus level between obese and lean subjects were clustered using complete linkage method. Values indicate abundance of microbiota at the genus level. **(B)** Average taxonomy abundance at genus level were represented as stacked bar plot to of lean and obese subject of each country. *x*-axis depicts percentage of abundance of each taxonomy. Dark and light blue represent obese and lean subjects, respectively. Red, blue, green, and brown represent individuals from Chile, United States, France, and Saudi Arabia, respectively.

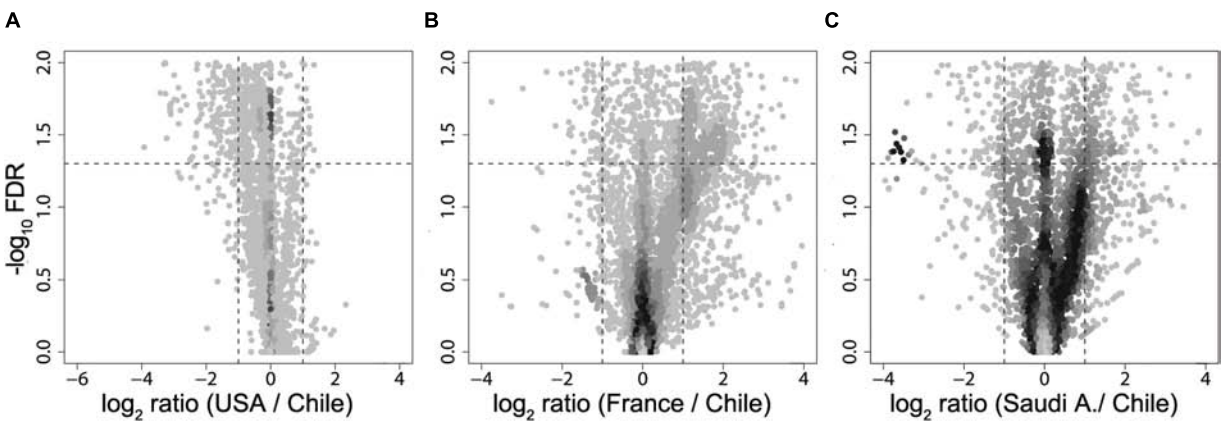


FIGURE 3 | Volcano plot of KEGG Orthology abundances obtained from metagenomic simulation. Fold change of United States **(A)**, France **(B)**, and Saudi Arabia **(C)** of KO abundance with regard to Chile are plotted against $-\log_{10}$ of FDR. Dotted lines in the *x*-axis denotes a two-fold change, while dotted lines in the *y*-axis delimit a FDR value of 0.05.

sequencing in order to dissect the contribution of metabolic functional changes and microbiota compositions. Additionally, we compared the results obtained from the Chilean cohort with similar studies carried out from patients in Italy (Tremaroli et al., 2015) and Denmark (Palleja et al., 2016), described in **Table 1**. All raw data were analyzed using the Metagenomic SOAP v2 from Microbiome Helper (Comeau et al., 2017) as described in "Materials and Methods." First, we compared the functional composition of obese subjects from Chile ($n = 6$), Italy ($n = 7$) and Denmark ($n = 7$) before any medical intervention utilizing CCA, Adonis tests and pairwise Spearman rank correlation. Initially, we performed an unstratified analysis, which corresponds to the analysis of metabolic pathway enrichment without taking into account microbial composition, and observed no statistically significant differences between the three countries using CCA analysis and Adonis tests (p-value > 0.05), in which the variance plot displayed little dispersion between the samples from different countries (**Figure 4A**). Furthermore, the pairwise Spearman rank correlation showed overall good correlation between the samples of all datasets (**Figure 4B**). Hence, in contrast to PICRUSt prediction, these results showed that the overall microbial functionality had no differences between obese subjects belonging to different world regions. Subsequently, we performed a stratified data analysis, a type of analysis that distinguishes the functional contribution of each bacterial species to the overall metabolic pathways. Interestingly, stratified data, unlike

unstratified analysis, showed significant dispersion between the countries (**Figure 4C**), and no correlation between datasets was observed (**Figure 4D**), indicating that individual microbial metagenomic contribution to overall functional pathways is different between countries, in agreement with functional inferences obtained by PICRUSt (**Figure 3**). Also, we use the Spearman rank correlation values obtained to compare statistically into-countries and cross-country associations in both unstratified and stratified data, finding no statistical differences between Spearman values (*p*-value FDR adjusted of 0.75 and 1, respectively). Utilizing linear discriminant analysis (LDA) to quantify phylogenetic and functional diversity, pathways over 1.5 LDA score were found to be differentially enriched between countries in both stratified and unstratified approaches (**Figure 5**). In addition, we validated the PICRUSt functional prediction comparing KEGG Orthology obtained with HUMAnN2 output from metagenomic data using Pearson and pairwise Spearman rank Correlation, with both showing linear dependence between 3590 common KEGG Orthology families (**Supplementary Figure S4** and **Supplementary Table S5**), which proved the relationship between *in silico* predictions and our biological findings as was previously demonstrated by PICRUSt authors (Langille et al., 2013). Altogether, these results suggested that microbial contribution to functional pathways was country-specific, indicating that there was redundancy in the functions of these distinct microbial species found in obese subjects from different countries, in which different microbiota components contribute to obtain similar metabolic outcomes.

Furthermore, microbial taxonomy abundance data obtained from shotgun DNA sequencing showed that Chilean microbiota composition from obese subjects was different compared to Italian and Danish cohorts (Tremaroli et al., 2015; Palleja et al., 2016). CCA analysis and Adonis tests showed significant differences between the microbiota of analyzed subjects at species level (**Supplementary Figure S5A**), which strengthens our previous observations obtained from the 16S rDNA sequencing taxonomic comparison at genus level (**Figure 1B**). In addition, LDA analysis showed differences in species abundance between the three microbial datasets species with and LDA scores higher than 2, reinforcing the observation that the microbiota under dysbiosis has different compositions specific to each geographical location (**Supplementary Figure S5B**).

Gut Microbiota Functionality Changes Following Bariatric Surgery in Chilean Subjects

Several studies have shown that the regulation of energy and fat storage is influenced by intestinal microbes, and this composition may contribute to obesity (Qin et al., 2012; Karlsson et al., 2013). In this regard, it has been reported that human gut microbiota composition varies between obese patients who underwent different types of surgical intervention (Zhang et al., 2009; Furet et al., 2010; Schloss, 2016; Ilhan et al., 2017; Medina et al., 2017; Murphy et al., 2017). Moreover, several of these studies describe functional changes after treatment and its prevalence across time (Tremaroli et al., 2015; Palleja et al., 2016). More specifically, they report an increase in Proteobacteria abundance and facultative anaerobes, such as *E. coli*, and an enrichment in microbial pathways involving aerobic respiration, glutathione and gamma-aminobutyric acid metabolism. Similarly, in a previous study from our laboratory using Chilean obese patients, we observed significant changes in the human gut microbiota, and these changes were specific to the type of surgery performed (Medina et al., 2017). Here, utilizing the same cohort, we performed shotgun DNA sequencing of stool samples before and after surgical intervention in order to analyze changes in taxonomic composition (Medina et al., 2017). Although no statistical testing was performed due to insufficient cohort number (n = 2 for each group), our results suggested that the two kinds of bariatric surgery mediated different rearrangements of functional pathways. We found stratified and unstratified functional changes of several folds, 6 months following either Roux-en-Y gastric bypass (RYGB) or sleeve gastrectomy (SG) interventions (**Figure 6**). More specifically, unstratified functional changes were more pronounced in RYGB–treated patients compared to SG-treated patients, represented by higher data variance with an increase in $\log_2$ fold change (**Figures 6B,D**). Functional changes in unstratified RYGB data suggested an increase in pathways related with acetyl-CoA biosynthesis, trehalose degradation, GABA shunt and phospholipid remodeling. Contrastingly, changes observed in SG intervention included an increase in metabolic pathways such as fatty acid β-oxidation, TCA cycle and acetyl-CoA biosynthesis (**Supplementary Table S4**). Top stratified functional changes after RYGB were mainly driven by *A. muciniphila*, *E. coli*, *Bacteroides vulgatus*, *Eubacterium siraeum* and *Streptococcus salivarius*, while in SG functional changes were driven by *Bacteroides cellulosilyticus*, *S. salivarius*, *Eubacterium eligens*, *Lactococcus lactis*, *Alistipes finegoldii, E. coli* and *A. muciniphila* (**Supplementary Table S4**). Altogether, these results suggested that bariatric surgery in Chilean patients also caused functional rearrangement in microbiota, in concordance to previous metagenomic studies (Graessler et al., 2013; Tremaroli et al., 2015; Palleja et al., 2016), and these changes appeared to be specific to the surgery performed.

DISCUSSION

Obesity is a world-wide health problem whose global prevalence has increased at an accelerated rate since 1980 (Stevens et al., 2012). It has been previously described that the composition and functionality of human gut microbiota plays a crucial role in mediating both disease and recovery following medical intervention (Turnbaugh et al., 2006; Tremaroli and Backhed, 2012; Karlsson et al., 2013; Le Chatelier et al., 2013; Damms-Machado et al., 2015). This study compared the gut microbiota compositions of obese subjects from Chile with previously published data from the United States, France, and Saudi Arabia using 16S rDNA Illuminia MiSeq, as well as from Italy and Denmark using Illumina HiSeq shotgun metagenomics, and found that the microbiota composition differs significantly between obese and lean subjects from these different countries, showing country-specific clustering. It has previously been

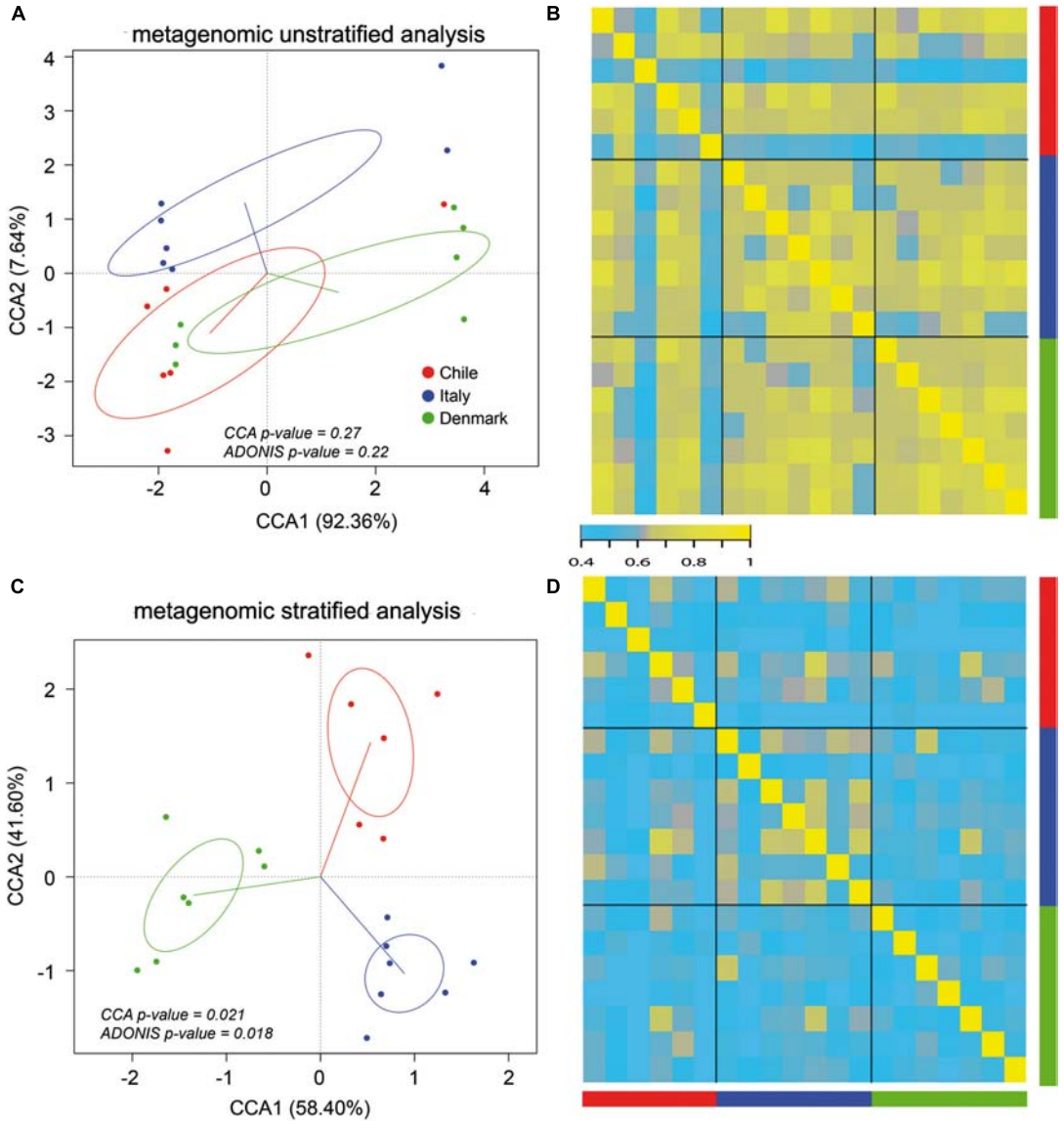


FIGURE 4 | Stratified and unstratified functional profiles of obese gut microbiota from three different countries. Unstratified **(A,B)** and stratified **(C,D)** analyses of obese gut microbiota abundance for Chile (red), Italy (blue) and Denmark (green). Scatterplots **(A,C)** represent CCA and Adonis test analysis, and heatmap **(B,D)** represent pairwise Spearman rank correlation of functional microbial abundance between countries.

described that a number of factors can modulate the composition of intestinal microbiota, one of which is geographic origin (reviewed by Rojo et al., 2017). A previous study compared the microbiota composition of a group of healthy subjects from France with Saudi Arabia by 16S rDNA sequencing and suggested that the dietary habits of different regions or countries can generate changes in intestinal microbiota (Yasir et al., 2015). This idea was reinforced by other studies suggesting that different types of diets directly change microbiota composition (Wu et al., 2011; Fava et al., 2013). In fact, mice subjected to obesogenic diets displayed significant functional and taxonomic modifications in their gut microbiota (Tran et al., 2019). Other studies show that genetics play a secondary role compared to environmental factors (Gupta et al., 2017; Jones et al., 2018; Rothschild et al., 2018). However, twin studies revealed more highly correlated microbiota compositions between monozygotic twins than dizygotic twins (Goodrich et al., 2014). Given subjects belonging to close location have less genetic variance than subjects from distant regions, this may also contribute to the differences in microbiota composition observed between geographical origins.

Our work attempted to indicate the need for the global study of human microbiota for a proper assessment of the microbial contribution in respect to geographic distribution. In this context, we compared the taxonomic gut microbiota abundance of lean and obese at genus level using multivariate variance

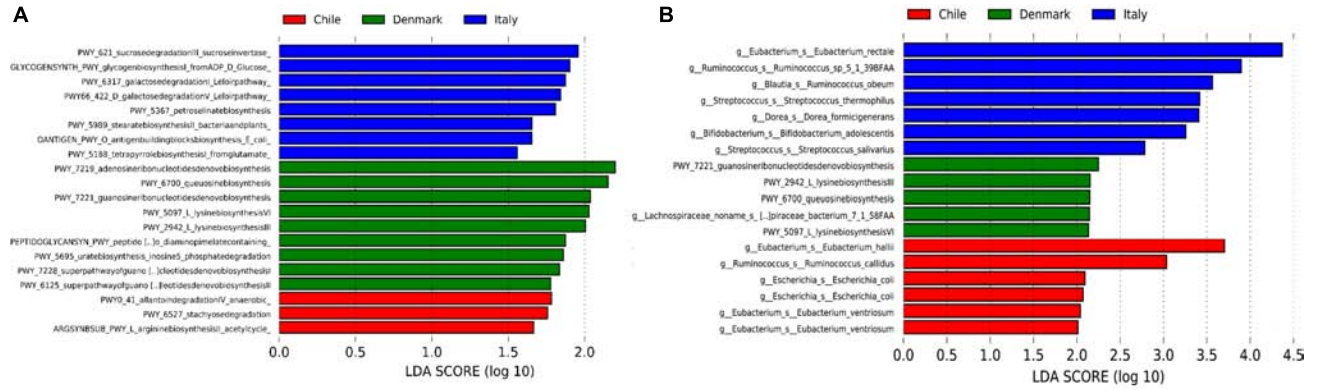


FIGURE 5 | Linear discriminant analysis (LDA) effect size (LEfSe) of metabolic pathways. Unstratified **(A)** and stratified **(B)** functional enrichment analyses from microbiome of obese subjects belonging to Chile (red, $n = 6$), Denmark (green, $n = 7$) and Italy (blue, $n = 7$). Pathways with LDA scores higher than 1.5 **(A)** or 2 **(B)** are shown. For stratified functional enrichment **(B)**, unmapped/unintegrated reads are denoted as the bacterial taxa in which the pathways belong to.

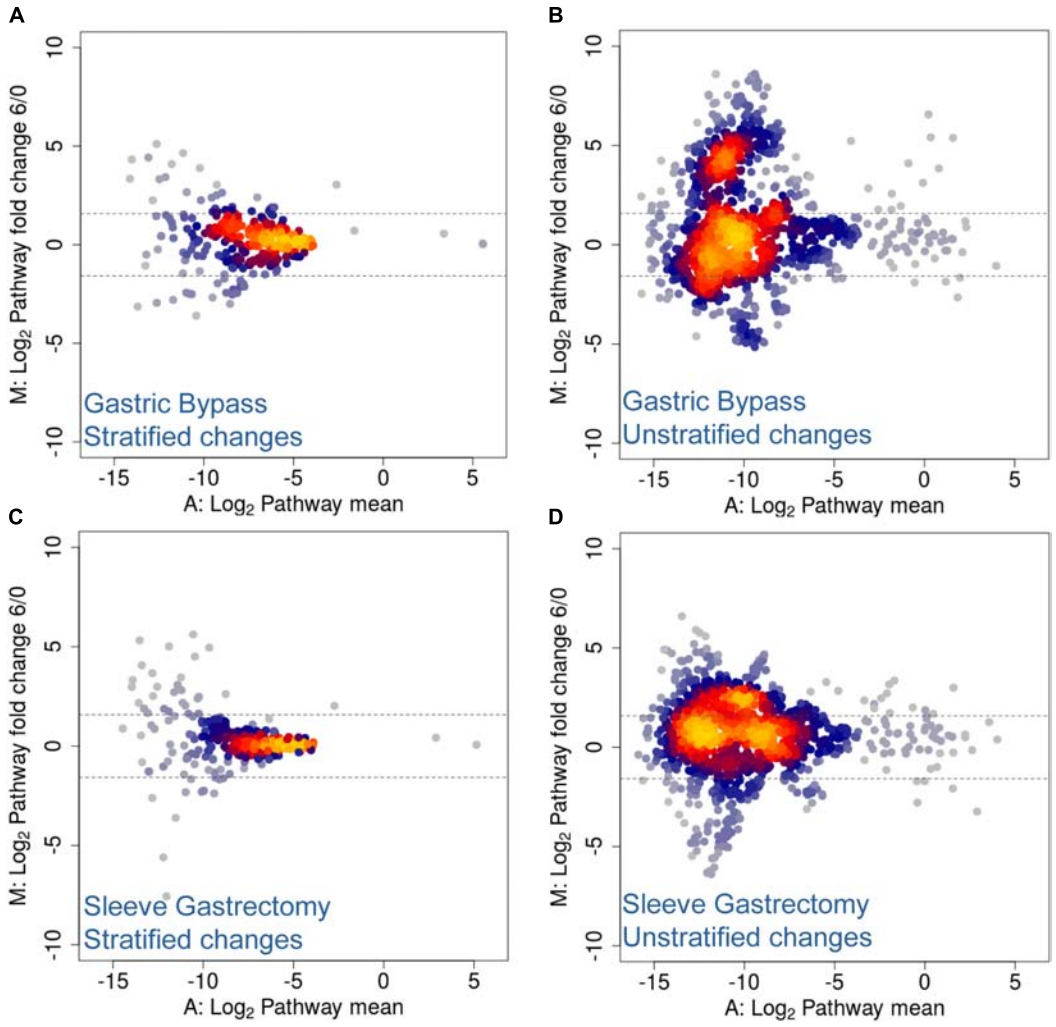


FIGURE 6 | MA Plots for stratified and unstratified functional changes after bariatric surgery in Chilean patients. **(A,B)** Gastric bypass and **(C,D)** sleeve gastrectomy stratified **(A,C)** and unstratified **(B,D)** functional changes 6 months post-surgery. Values are expressed as mean (x-axis) and fold change (y-axis), and represented as $\log_2$ values. Dotted lines denote a two-fold change.

analysis and non-parametric mean comparison. Although, using Wilcoxon rank test, we found different statistical changes in France and Chile gut microbiota, this was, however, not the case for United States and Saudi Arabia. Our results suggest a country-specific signature for both lean and obese microbiota. Nevertheless, we cannot discard the possibility that our observations may be influenced by technical and methodological factors. First, it is known that non-parametric tests have less chance of detecting a true effect where one exists (Whitley and Ball, 2002). Hence, utilizing a different statistical approach to analyze differences in variance (CAA and Adonis test) between lean and obese subjects in France, Chile, and Saudi Arabia, the subsequent results support the notion that obese versus lean microbiota is different between countries. Second, we cannot discard that this observation may be relatively influenced by the methodology differences between studies and also may be confounded by large interpersonal variation. A meta-analysis of the association between differences in the microbiome and obesity status made between 10 published studies found that is difficult to classify a subject as obese by its microbiota composition, suggesting the possibility that each individual has their own bacterial signatures (Sze and Schloss, 2016). Another meta-analysis that compared lean and obese subject from five published studies also found that signatures of obesity were not consistent between studies, as just one of the 5 studies analyzed showed diversity differences between lean and obese subject (Walters et al., 2014). Nevertheless, the authors declared that it was possible to assess differences between lean and obese subjects when they changed the method to identify OTUs (from "close reference" to "*pick de novo*") or used supervised learning tools to categorize subjects according to lean and obese states with considerable accuracy (Walters et al., 2014). Altogether, these observations suggest that geographical differences between lean and obese subjects requires additional validation using an international cohort that not only takes into account confounding factors but maintain the same experimental design.

Our study performed a functional inference analysis based on the data obtained by 16S rDNA sequencing and observed significant differences in KEGG Orthology enrichment in samples obtained from France, United States, and Saudi Arabia compared to Chile. These results suggested the microbial differences observed were associated with changes in functional pathways, however, this kind of analysis was unable to differentiate the microbial contribution of each microorganism species to the identified functional pathways. Therefore, a more detailed comparison of metagenomic DNA sequencing data was used, which was obtained using the HiSeq platform to analyze stool samples of obese patients from Chile, Italy and Denmark. First, we performed an unstratified analysis, a type of metabolic pathway enrichment analysis without taking into account microbial contribution, and revealed that the functionality in obesity context did not show statistical differences (in both CCA and Adonis testing) between patients from Chile, Italy and Denmark, suggesting that gut microbiota metabolic pathways do not change according to geographic origin. Conversely, using a stratified approach that could identify specific microbial composition and their individual contribution to metabolic pathways, significant differences between countries were identified, in agreement with our initial PICRUSt metabolic inference, because both methods takes in account taxonomic composition to assess metabolic pathways abundance. Altogether, our observations indicate while overall metabolic functions do not change, microbial functional contribution to obesity are specific to each country.

Although the lack of differences observed in metagenomic unstratified analysis may be due to small cohort numbers, our comparisons suggest that the gut microbiota of obese patients from different countries have functional convergence, where the same essential metabolic functions are carried out by related and unrelated bacterial species. Nevertheless, it is not surprising that bacterial compositions of obese individuals differ between counties due to both genetic and environmental factors, as previously described by others (reviewed by Gupta et al., 2017), however, our observations relating to convergent metabolic functions provide new insights into how the gut microbiota shape common disease phenotypes. The concept of intestinal functional redundancy has been previously explored, in which taxonomic diversity appears to be irrelevant for the inference of functional traits (reviewed by Moya and Ferrer, 2016; Rojo et al., 2017). One such example was demonstrated by the Human Microbiome Project Consortium (Huttenhower et al., 2012), where the authors observed the microbial metabolism to remain constant across individuals over time despite high variability in composition. One possible explanation of microbial functional redundancy can be due to the evolutionary convergence of unrelated taxa, in which variable combinations of species from different phyla can at least partially fulfill the metabolic functions of another, resulting in different species of bacteria behaving similarly (Moya and Ferrer, 2016; Rojo et al., 2017).

Bariatric surgeries have become a frequent choice of treatment for obesity patients over the last years, being SG the most frequent procedure in the world (Angrisani et al., 2017; reviewed by ASMBS, 2018). Several studies have shown that the intestinal microbiota composition changes following surgical intervention and these changes vary between patients who have undergone different kind of surgical intervention such as RYGB, SG and laparoscopic adjustable gastric banding (LAGB) (Tremaroli et al., 2015; Palleja et al., 2016; Ilhan et al., 2017; Medina et al., 2017). One of the limitations of our metagenomic study after bariatric surgery is our small cohort number (n = 2 for each group), and therefore, further studies with larger cohort are required to confirm that different treatments mediate differentially metagenomic rearrangements. However, this study suggest that the microbial functionality of patients changed 6 months following surgical treatment. In agreement with previous studies, it also observed specificity in the changes regarding to the type of surgery performed, where functional changes were mainly mediated by *A. muciniphila*, *E. coli*, *B. vulgatus*, *E. siraeum* and *S. salivarius* after RYGB, and by *B. cellulosilyticus*, *S. salivarius*, *E. eligens*, *L. lactis*, *A. finegoldii*, *E. coli* and *A. muciniphila* species after SG. Although it is impossible to determine from this study whether the microbiota composition changes were a consequence of dietary changes or weight modification, our results not only

hint at a microbiota signature for different bariatric surgeries, but also suggest the need for further microbiota meta-analyses at world-wide levels to study metabolic disorders such as obesity.

A common key factor in obesity is the low abundance of *A. muciniphila* (Schneeberger et al., 2015; Palleja et al., 2016; Medina et al., 2017; Seck et al., 2018), a mucin-degrading bacterium that resides in the human gut mucus layer and whose abundance in healthy subjects represents 3–5% of the residential microbial community (Derrien et al., 2004). Interestingly, it has been shown that *A. muciniphila* prevents inflammation and adipose tissue alterations in mice (Schneeberger et al., 2015). The administration of *A. muciniphila* grown under mucin-depleted conditions is effective in reducing obesity and improve intestinal barrier integrity in obese mice (Shin et al., 2019), controls fat mass storage and glucose homeostasis in obese and type 2 diabetic mice (Everard et al., 2013), and it has been previously described that bariatric surgery improves its abundance (Damms-Machado et al., 2015; Tremaroli et al., 2015; Palleja et al., 2016; Medina et al., 2017). Furthermore, overweight and obese individuals with higher *A. muciniphila* abundance is associated with a healthier metabolic status compared with lower abundance (Dao et al., 2016). Therefore, the enrichment of *A. muciniphila* provides a possible therapy in the treatment of obesity, and this possibility has been explored in a recent study (Sheng et al., 2018). Here, we identified *A. muciniphila* as one of the bacteria that drove the changes in metabolic pathways after surgical intervention of Chilean obese patients, an important observation considering that gut microbiota of healthy Chilean subjects has high abundance of *Verrucomicrobia* bacteria, including *A. muciniphila* (Fujio-Vejar et al., 2017). Further studies to understand the underlying mechanisms involving *A. muciniphila* as a target of bacto-therapy to treat obesity are required.

Our results, together with previously published studies, highlight the need to consider region-specific analysis of the gut microbiota in order to fully understand the bacterial basis for the development of such diseases as obesity and the response to surgical and non-surgical treatment, opening to the possibility that probiotic development to treat different kinds of dysbiosis should be country-specific.

CONCLUSION

This study identified significant differences in the human gut microbiota of obese patients from around the world, and found that functional dissimilarities were mediated by differences in taxonomic microbiota composition, which were region-specific, rather than alterations in metabolic pathways. This indicates the presence of functional metabolic redundancy between the microbiota of obese patients despite the bacterial differences and geographic origin. Furthermore, functional changes in gut microbiota following bariatric surgery were observed to be specific to the type of treatment received, providing new insights into the role of the gut microbiome in treatment strategies.

ETHICS STATEMENT

All experiments were conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Faculty of Medicine, Pontifical Catholic University of Chile (n° 15–337). DNA samples used here for metagenomic belongs from previous published studies (Medina et al., 2017). The 16S and metagenomic raw data sequences used belongs from previous published studies (Tremaroli et al., 2015; Yasir et al., 2015; Palleja et al., 2016; Ilhan et al., 2017; Medina et al., 2017; Thomson et al., 2019).

AUTHOR CONTRIBUTIONS

DM conceived, designed, performed the comparisons, analyzed the data, and wrote the manuscript. TL wrote and edited the figures and manuscript. PT contributed to the discussion. AA provided the scripts to do the multivariate data analysis. VP-B and AM critically read and edited the manuscript.

FUNDING

This work was supported by CONICYT-Chile through the FONDECYT [n° 3160525] (DM) and by Universidad San Sebastián, and by grants to AM from the Spanish Ministry of Economy and Competitiveness (projects SAF2012-31187, SAF2013-49788-EXP, SAF2015-65878-R), Carlos III Health Institute (Projects PIE14/00045 and AC15/00022), Generalitat Valenciana (Project PrometeoII/2014/065 and Prometeo/2018/A/133), Asociación Española Contra el Cancer (Project AECC 2017-1485) and co-financed by the European Regional Development Fund (ERDF).

ACKNOWLEDGMENTS

We thank Daniel Garrido from Pontificia Universidad Católica de Chile for kindly providing the DNA samples belonging to a previous study (Medina et al., 2017) to be used here for shotgun DNA sequencing. We also thank the Departamento de Ingeniería Química y Bioprocesos of the Pontificia Universidad Católica de Chile for their contribution.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.02346/full#supplementary-material

REFERENCES

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

Angelakis, E., Armougom, F., Million, M., and Raoult, D. (2012). The relationship between gut microbiota and weight gain in humans. *Future Microbiol.* 7, 91–109. doi: 10.2217/fmb.11.142

Angrisani, L., Santonicola, A., Iovino, P., Vitiello, A., Zundel, N., Buchwald, H., et al. (2017). Bariatric surgery and endoluminal procedures: IFSO worldwide survey 2014. *Obes. Surg.* 27, 2279–2289. doi: 10.1007/s11695-017-2666-x

ASMBS (2018). American Society for Metabolic and Bariatric Surgery: Estimate of Bariatric Surgery Numbers, 2011–2016. Available at: https://asmbs.org/resources/estimate-of-bariatric-surgery-numbers.

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

Bolstad, B. (2019). *Preprocesscore: A Collection of Pre-Processing Functions*. Avalaible at: https://github.com/bmbolstad/preprocessCore.

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. Publ. Gr.* 7, 335–336. doi: 10.1038/nmeth.f.303

Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M., and Owen, L. J. (2015). Dysbiosis of the gut microbiota in disease. *Microb. Ecol. Health Dis.* 26:26191. doi: 10.3402/mehd.v26.26191

Chilton, S. N. (2014). The effects of diet and the microbiome on reproduction and longevity: a comparative review across 5 continents. *J. Nutr. Food Sci.* 5:364. doi: 10.4172/2155-9600.1000364

Comeau, A. M., Douglas, G. M., and Langille, M. G. (2017). Microbiome helper: a custom and streamlined workflow for microbiome. *mSystems* 3:2. doi: 10.1128/mSystems.00127-16

Damms-Machado, A., Mitra, S., Schollenberger, A. E., Kramer, K. M., Meile, T., Königsrainer, A., et al. (2015). Effects of surgical and dietary weight loss therapy for obesity on gut microbiota composition and nutrient absorption. *Biomed Res. Int.* 2015:806248. doi: 10.1155/2015/806248

Dao, M. C., Everard, A., Aron-Wisnewsky, J., Sokolovska, N., Prifti, E., Verger, E. O., et al. (2016). Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut* 65, 426–436. doi: 10.1136/gutjnl-2014-308778

Derrien, M., Vaughan, E. E., Plugge, C. M., and de Vos, W. M. (2004). Akkermansia municiphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int. J. Syst. Evol. Microbiol.* 54(Pt 5), 1469–1476. doi: 10.1099/ijs.0.02873-0

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.03006-5

Eldar, S., Heneghan, H. M., Brethauer, S. A., and Schauer, P. R. (2011). Bariatric surgery for treatment of obesity. *Int. J. Obes.* 35(Suppl. 3), S16–S21. doi: 10.1038/ijo.2011.142

Everard, A., Belzer, C., Geurts, L., Ouwerkerk, J. P., Druart, C., Bindels, L. B., et al. (2013). Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 9066–9071. doi: 10.1073/pnas.1219451110

Fava, F., Gitau, R., Griffin, B. A., Gibson, G. R., Tuohy, K. M., and Lovegrove, J. A. (2013). The type and quantity of dietary fat and carbohydrate alter faecal microbiome and short-chain fatty acid excretion in a metabolic syndrome "at-risk" population. *Int. J. Obes.* 37, 216–223. doi: 10.1038/ijo.2012.33

Fei, T., Zhang, T., Shi, W., and Yu, T. (2018). Mitigating the adverse impact of batch effects in sample pattern detection. *Bioinformatics* 34, 2634–2641. doi: 10.1093/bioinformatics/bty117

First, Y., and Job, P. (2011). GNU parallel: the command-line power tool. *USENIX Mag.* 3, 42–47. doi: 10.5281/zenodo.16303

Fujio-Vejar, S., Vasquez, Y., Morales, P., Magne, F., Vera-Wolf, P., Ugalde, J. A., et al. (2017). The gut microbiota of healthy chilean subjects reveals a high abundance of the phylum verrucomicrobia. *Front. Microbiol.* 8:1221. doi: 10.3389/fmicb.2017.01221

Furet, J., Kong, L., Tap, J., Poitou, C., Basdevant, A., Bouillot, J. L., et al. (2010). Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss. *Diabetes Metab. Res. Rev.* 59, 3049–3057. doi: 10.2337/db10-0253

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

Graessler, J., Qin, Y., Zhong, H., Zhang, J., Licinio, J., Wong, M. L., et al. (2013). Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters. *Pharmacogenomics J.* 13, 514–522. doi: 10.1038/tpj.2012.43

Gumbs, A. A., Gagner, M., Dakin, G., and Pomp, A. (2007). Sleeve gastrectomy for morbid obesity. *Obes. Surg.* 17, 962–969. doi: 10.1007/s11695-007-9151-x

Guo, Y., Zhao, S., Su, P. F., Li, C. I., Ye, F., Flynn, C. R., et al. (2014). Statistical strategies for microRNAseq batch effect reduction. *Transl. Cancer Res.* 3, 260–265. doi: 10.3978/j.issn.2218-676X.2014.06.05

Gupta, V. K., Paul, S., and Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front. Microbiol.* 8:1162. doi: 10.3389/fmicb.2017.01162

Hall, A. B., Tolonen, A. C., and Xavier, R. J. (2017). Human genetic variation and the gut microbiome in disease. *Nat. Rev. Genet.* 18, 690–699. doi: 10.1038/nrg.2017.63

Haro, C., Rangel-Zúñiga, O. A., Alcalá-Díaz, J. F., Gómez-Delgado, F., Pérez-Martínez, P., Delgado-Lista, J., et al. (2016). Intestinal microbiota is influenced by gender and body mass index. *PLoS One* 2:11. doi: 10.1371/journal.pone.0154090

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

Ilhan, Z. E., Dibaise, J. K., Isern, N. G., Hoyt, D. W., Marcus, A. K., Kang, D. W., et al. (2017). Distinctive microbiomes and metabolites linked with weight loss after gastric bypass, but not gastric banding. *ISME J.* 11, 2047–2058. doi: 10.1038/ismej.2017.71

Jones, R. B., Zhu, X., Moan, E., Murff, H. J., Ness, R. M., Seidner, D. L., et al. (2018). Inter-niche and inter-individual variation in gut microbial community assessment using stool, rectal swab, and mucosal samples. *Sci. Rep.* 8:4139. doi: 10.1038/s41598-018-22408-4

Karlsson, F., Tremaroli, V., Nielsen, J., and Backhed, F. (2013). Assessing the human gut microbiota in metabolic diseases. *Diabetes Metab. Res. Rev.* 62, 3341–3349. doi: 10.2337/db13-0844

Kasai, C., Sugimoto, K., Moritani, I., Tanaka, J., Oya, Y., Inoue, H., et al. (2015). Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC Gastroenterol.* 15:100. doi: 10.1186/s12876-015-0330-2

Khan, M. J., Gerasimidis, K., Edwards, C. A., and Shaikh, M. G. (2016). Role of gut microbiota in the aetiology of obesity: proposed mechanisms and review of the literature. *J. Obes.* 2016, 1–27. doi: 10.1155/2016/7353642

Kolde, R., Franzosa, E. A., Rahnavard, G., Hall, A. B., Vlamakis, H., Stevens, C., et al. (2018). Host genetic variation and its microbiome interactions within the human microbiome project. *Genome Med.* 10:6. doi: 10.1186/s13073-018-0515-8

Kong, L. C., Tap, J., Aron-Wisnewsky, J., Pelloux, V., Basdevant, A., Bouillot, J. L., et al. (2013). Gut microbiota after gastric bypass in human obesity: increased richness and associations of bacterial genera with adipose tissue genes. *Am. Soc. Nutr.* 98, 16–24. doi: 10.3945/ajcn.113.058743

Langille, M. G. I., 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

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

Le Chatelier, E., 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. doi: 10.1038/nature12506

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/nature4441021a

Li, J. V., Ashrafian, H., Bueter, M., Kinross, J., Sands, C., and le Roux, C. W. (2011). Metabolic surgery profoundly influences gut microbial-host metabolic cross-talk. *Gut* 60, 1214–1223. doi: 10.1136/gut.2010.234708

Louis, S., Tappu, R. M., Damms-Machado, A., Huson, D. H., and Bischoff, S. C. (2016). Characterization of the gut microbial community of obese patients following a weight-loss intervention using whole metagenome shotgun sequencing. *PLoS One* 11:1–18. doi: 10.1371/journal.pone.0149564

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, 220–230. doi: 10.1038/nature11550

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, 610–618. doi: 10.1038/ismej.2011.139

McIver, L. J., Abu-Ali, G., Franzosa, E. A., Schwager, R., Morgan, X. C., Waldron, L. et al. (2018). bioBakery: a meta'omic analysis environment. *Bioinformatics* 34, 1235–1237. doi: 10.1093/bioinformatics/btx754

Medina, D. A., Pedreros, J. P., Turiel, D., Quezada, N., Pimentel, F., Escalona, A., et al. (2017). Distinct patterns in the gut microbiota after surgical or medical therapy in obese patients. *PeerJ* 5:e3443. doi: 10.7717/peerj.3443

Million, M., Maraninch, M., Henry, M., Armougom, F., Richet, H., Carrieri, P., et al. (2012). Obesity-associated gut microbiota is enriched in Lactobacillus reuteri and depleted in *Bifidobacterium animalis* and *Methanobrevibacter smithii*. *Int. J. Obes.* 36, 817–825. doi: 10.1038/ijo.2011.153

Moya, A., and Ferrer, M. (2016). Functional redundancy-induced stability of gut microbiota subjected to disturbance. *Trends Microbiol.* 24, 402–413. doi: 10.1016/j.tim.2016.02.002

Murphy, R., Tsai, P., Jüllig, M., Liu, A., Plank, L., and Booth, M. (2017). Differential changes in gut microbiota after gastric bypass and sleeve gastrectomy bariatric surgery vary according to diabetes remission. *Obes. Surg.* 27, 917–925. doi: 10.1007/s11695-016-2399-2

Navas-Molina, J. A., Peralta-Sanchez, J. M., Gonza, Lez, A., McMurdie, P. J., Vázquez-Baeza, Y., et al. (2013). Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.* 531, 371–444. doi: 10.1016/B978-0-12-407863-5.00019-8

NHLBI Expert Panel (1998). Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults - the evidence report. *Obes. Res.* 6(Suppl. 2), 51S–209S.

Nishijima, S., Suda, W., Oshima, K., Kim, S.-W., Hirose, Y., Morita, H., et al. (2016). The gut microbiome of healthy Japanese and its microbial and functional uniqueness. *DNA Res.* 23, 125–133. doi: 10.1093/dnares/dsw002

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2018). *Vegan: Community Ecology Package*. Available at: https://cran.r-project.org/package=vegan.

Palleja, A., Kashani, A., Allin, K. H., Nielsen, T., Zhang, C., Li, Y., et al. (2016). Roux-en-Y gastric bypass surgery of morbidly obese patients induces swift and persistent changes of the individual gut microbiota. *Genome Med.* 8, 67. doi: 10.1186/s13073-016-0312-1

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

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. doi: 10.1038/nature11450

R Core Team (2013). *R Core Team. R A Lang. Environ. Stat. Comput. R Found. Stat. Comput*. Vienna: R Core Team.

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

Rojo, D., Méndez-García, C., Raczkowska, B. A., Bargiela, R., Moya, A., Ferrer, M., et al. (2017). Exploring the human microbiome from multiple perspectives: factors altering its composition and function. *FEMS Microbiol. Rev.* 41, 453–478. doi: 10.1093/femsre/fuw046

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, 210–215. doi: 10.1038/nature25973

Schloss, P. D. (2016). Looking for a signal in the noise: revisiting obesity and the. 7, 1–9. doi: 10.1128/mBio.01018-16.Editor

Schneeberger, M., Everard, A., Gómez-Valadés, A. G., Matamoros, S., Ramírez, S., Delzenne, N. M., et al. (2015). Akkermansia muciniphila inversely correlates with the onset of inflammation, altered adipose tissue metabolism and metabolic disorders during obesity in mice. *Sci. Rep.* 5:16643. doi: 10.1038/srep16643

Schwalb, B., Tresch, A., Torkler, P., Duemcke, S., Demel, C., Ripley, B., et al. (2011). *LSD: Lots of Superior Depictions*. Available at: http://cran.r-project.org/web/packages/LSD/index.html

Seck, E. H., Senghor, B., Merhej, V., Bachar, D., Cadoret, F., Robert, C., et al. (2018). Salt in stools is associated with obesity, gut halophilic microbiota and Akkermansia muciniphila depletion in humans. *Int. J. Obes.* 43, 862–871. doi: 10.1038/s41366-018-0201-3

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. doi: 10.1186/gb-2011-12-6-r60

Shao, Y., Ding, R., Xu, B., Hua, R., Shen, Q., He, K., et al. (2016). Alterations of gut microbiota after roux-en-Y gastric bypass and sleeve gastrectomy in sprague-dawley rats. *Obes. Surg.* 27, 295–302. doi: 10.1007/s11695-016-2297-7

Sheng, L., Jena, P. K., Liu, H.-X., Hu, Y., Nagar, N., Bronner, D. N., et al. (2018). Obesity treatment by epigallocatechin-3-gallate–regulated bile acid signaling and its enriched *Akkermansia muciniphila*. *FASEB J.* doi: 10.1096/fj.201800370R [Epub ahead of print]

Shin, J., Noh, J.-R., Chang, D.-H., Kim, Y.-H., Kim, M. H., Lee, E. S., et al. (2019). Elucidation of akkermansia muciniphila probiotic traits driven by mucin depletion. *Front. Microbiol.* 10:1137. doi: 10.3389/fmicb.2019.01137

Sjöström, L. (2008). Bariatric surgery and reduction in morbidity and mortality: experiences from the SOS study. *Int. J. Obes.* 32(Suppl. 7), S93–S97. doi: 10.1038/ijo.2008.244

Smoot, T. M., Xu, P., Hilsenrath, P., Kuppersmith, N. C., and Singh, K. P. (2006). Gastric bypass surgery in the united states, 1998-2002. *Am. J. Public Health* 96, 1187–1189. doi: 10.2105/AJPH.2004.060129

Stevens, G. A., Singh, G. M., Lu, Y., Danaei, G., Lin, J. K., Finucane, M. M., et al. (2012). National, regional, and global trends in adult overweight and obesity prevalences. *Popul Health Metr.* 10:22. doi: 10.1186/1478-7954-10-22

Sun, Z., Chai, H. S., Wu, Y., White, W. M., Donkena, K. V., Klein, C. J., et al. (2011). Batch effect correction for genome-wide methylation data with illumina infinium platform. *BMC Med Genom.* 16:84. doi: 10.1186/1755-8794-4-84

Sze, M. A., and Schloss, P. D. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. *MBio* 23, 7. doi: 10.1128/mBio.01018-6

Thomson, P., Santibañez, R., Aguirre, C., Galgani, J., and Garrido, D. (2019). Short-term impact of sucralose consumption on the metabolic response and gut microbiome of healthy adults. *Br. J. Nutr.* 1, 1–23. doi: 10.1017/S0007114519001570

Tice, J. A., Karliner, L., Walsh, J., Petersen, A. J., and Feldman, M. D. (2008). Gastric banding or bypass? a systematic review comparing the two most popular bariatric procedures. *Am. J. Med.* 121, 885–893. doi: 10.1016/j.amjmed.2008.05.036

Tran, D. D., Nwokeabia, I. D., Purnell, S., Zafar, S. N., Ortega, G., Hughes, K., et al. (2016). Revision of roux-en-y gastric bypass for weight regain: a systematic review of techniques and outcomes. *Obes. Surg.* 26, 1627–1634. doi: 10.1007/s11695-016-2201-5

Tran, H. Q., Mills, R. H., Peters, N. V., Holder, M. K., de Vries, G. J., Knight, R. et al. (2019). Associations of the fecal microbial proteome composition and proneness to diet-induced obesity. *Mol. Cell. Proteomics.* 18, 1864–1879. doi: 10.1074/mcp.ra119.001623

Tremaroli, V., and Backhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature* 489, 242–249. doi: 10.1038/nature11552

Tremaroli, V., Karlsson, F., Werling, M., Ståhlman, M., Kovatcheva-Datchary, P., Olbers, T., et al. (2015). Roux-en-Y gastric bypass and vertical banded gastroplasty induce long-term changes on the human gut microbiome contributing to fat mass regulation. *Cell Metab.* 22, 228–238. doi: 10.1016/j.cmet.2015.07.009

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, 902–903. doi: 10.1038/nmeth.3589

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, 480–484. doi: 10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1131. doi: 10.1038/nature05414

Walters, W. A., Xu, Z., and Knight, R. (2014). Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett.* 588, 4223–4233. doi: 10.1016/j.febslet.2014.09.039

Whitley, E., and Ball, J. (2002). Statistics review 6: nonparametric methods. *Crit. Care* 6, 509–513. doi: 10.1186/cc1820

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 7 334, 105–108. doi: 10.1126/science.1208344

Yasir, M., Angelakis, E., Bibi, F., Azhar, E. I., Bachar, D., Lagier, J. C., et al. (2015). Comparison of the gut microbiota of people in france and saudi arabia. *Nutr Diabetes* 27:e153. doi: 10.1038/nutd.2015.3

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, 222–227. doi: 10.1038/nature11053

Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yeisoo, Y., et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. U. S. A.* 106, 2365–2370. doi: 10.1073/pnas.0812600106

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate illumina paired-end reAd mergeR. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593

Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 352, 565–569. doi: 10.1126/science.aad3369

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.

Copyright © 2019 Medina, Li, Thomson, Artacho, Pérez-Brocal and Moya. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

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: info@frontiersin.org | +41 21 510 17 00

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