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^{ε}_{ij} such that:

$$R^{\varepsilon}_{ij} \sim \begin{cases} Unif(0,\varepsilon) & with\ probability\ p_i \\ LN(\mu_i, \sigma_i^2) & with\ probability\ 1 - p_i \end{cases} \tag{1}$$

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:

$$LN(r\,|\,\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi r}} \exp\left[-\frac{(\log r - \mu)^2}{2\sigma^2}\right], \tag{2}$$

in which, by definition, $Y = \log R$ is normally distributed with mean μ and variance σ^2. Let $y_{ij} = \log r^{\varepsilon}_{ij}$, 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

$$\begin{aligned} \ell(\theta_i\,|\,y_{ij}, z_{ij}) = & \sum_{j=1}^{m} z_{ij} \log\left[\eta_i + (1-p_i)\phi(y_{ij};\, \mu_i, \sigma_i^2)\right] \\ & + \sum_{j=1}^{m} (1 - z_{ij}) \log(1 - p_i) \\ & + \sum_{j=1}^{m} (1 - z_{ij}) \log \phi(y_{ij};\, \mu_i, \sigma_i^2), \end{aligned} \tag{3}$$

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

$$H = -\sum_{j=1}^{N} a_j \ln a_j, \tag{4}$$

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

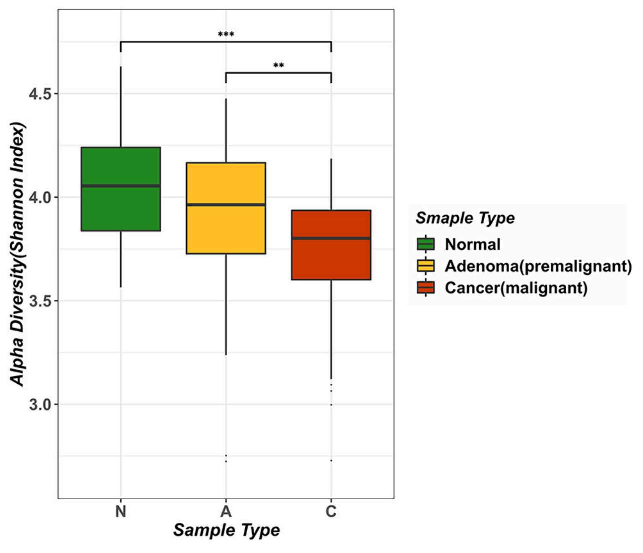

FIGURE 1 | Analysis of intestinal microbial diversity in different environments. The three colors in the figure indicate the microbial diversity in different states: green represents the healthy samples, yellow represents adenoma (precancerous lesion) growth in the intestine, and red represents a sample 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.

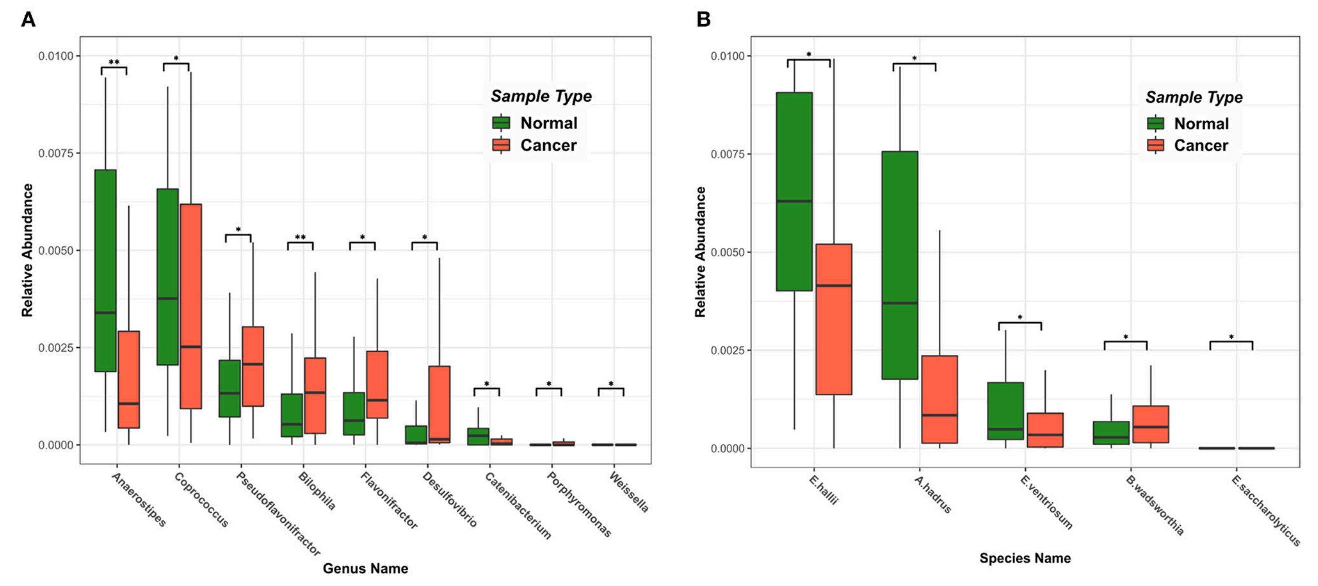

FIGURE 2 | Differences in intestinal microbiome at the genus and species level among samples of different states. Green represents a healthy sample, and red represents a colorectal cancer sample. **(A)** The top nine microbial genera with significant differences in abundance. **(B)** The top five microbial strains with significant differences in abundance.

logistic model, $P < 0.05$). The fitted logistic regression model was as follows:

$$P = \frac{\exp(-4.563d + 17.546)}{1 + \exp(-4.563d + 17.546)}, \tag{5}$$

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.

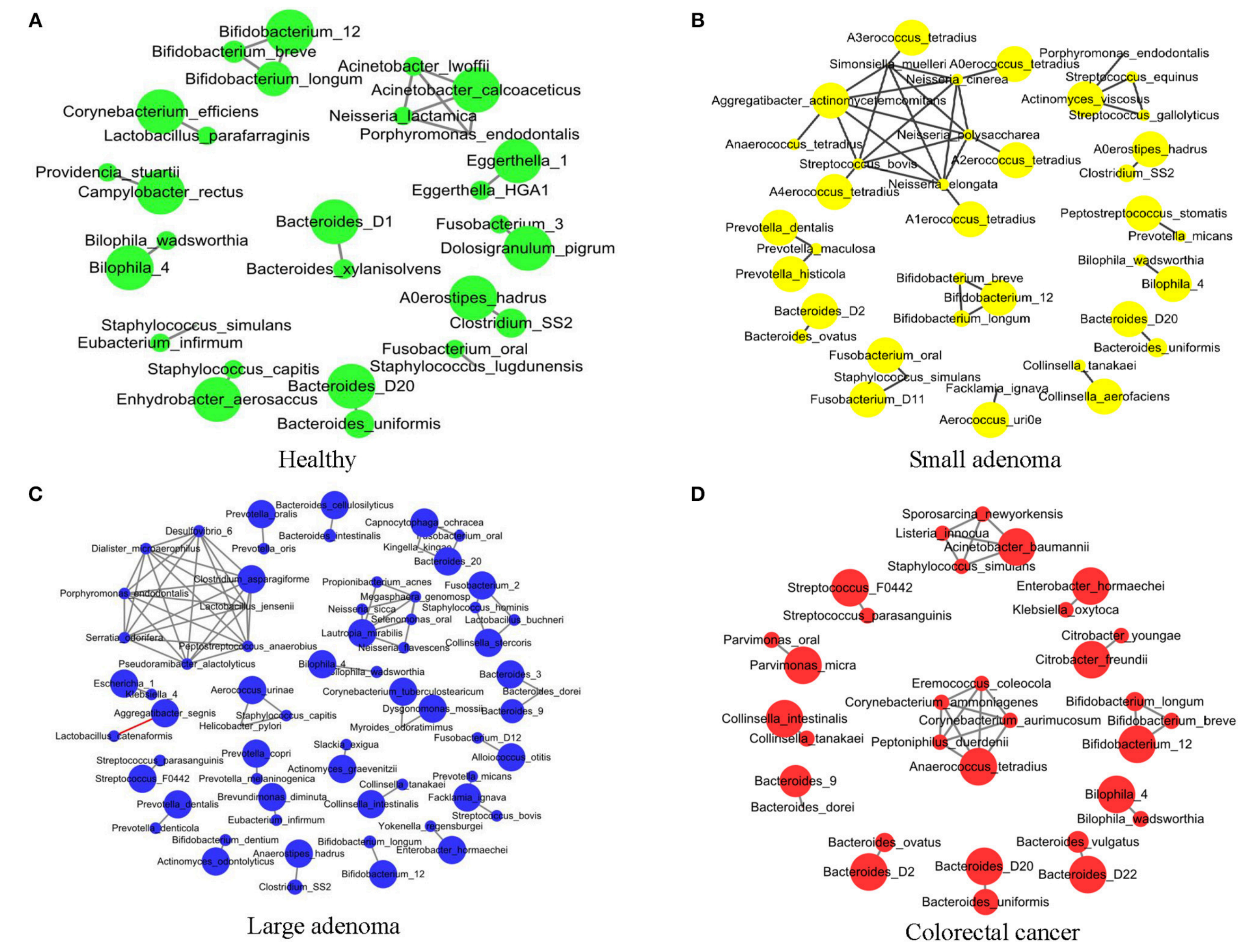

FIGURE 3 | The association network of intestinal microbiome in different states. Each circle represents the average relative abundance of a microbial species in that state. The higher the average relative abundance, the larger the area of the circle. The solid gray line between the circles indicates 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.