AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Host-Microbe Biology

Meta-analysis of 16S rRNA Microbial Data Identified Distinctive and Predictive Microbiota Dysbiosis in Colorectal Carcinoma Adjacent Tissue

Zongchao Mo,[a,b] Peide Huang,[b] Chao Yang,[b] Sihao Xiao,[b] Guojia Zhang,[b] Fei Ling,[a] Lin Li[b]

[a]School of Biology and Biological Engineering, South China University of Technology, Guangzhou, Guangdong, China
[b]BGI Genomics, BGI-Shenzhen, Shenzhen, China

Zongchao Mo and Peide Huang contributed equally to this work. Author order was determined in order of increasing seniority.

ABSTRACT As research focusing on the colorectal cancer fecal microbiome using shotgun sequencing continues, increasing evidence has supported correlations between colorectal carcinomas (CRCs) and fecal microbiome dysbiosis. However, large-scale on-site and off-site (surrounding adjacent) tissue microbiome characterization of CRC was underrepresented. Here, considering each taxon as a feature, we demonstrate a machine learning-based method to investigate tissue microbial differences among CRC, colorectal adenoma (CRA), and healthy control groups using 16S rRNA data sets retrieved from 15 studies. A total of 2,099 samples were included and analyzed in case-control comparisons. Multiple methods, including differential abundance analysis, random forest classification, cooccurrence network analysis, and Dirichlet multinomial mixture analysis, were conducted to investigate the microbial signatures. We showed that the dysbiosis of the off-site tissue of colonic cancer was distinctive and predictive. The AUCs (areas under the curve) were 80.7%, 96.0%, and 95.8% for CRC versus healthy control random forest models using stool, tissue, and adjacent tissue samples and 69.9%, 91.5%, and 89.5% for the corresponding CRA models, respectively. We also found that the microbiota ecologies of the surrounding adjacent tissues of CRC and CRA were similar to their on-site counterparts according to network analysis. Furthermore, based on the enterotyping of tissue samples, the cohort-specific microbial signature might be the crux in addressing classification generalization problems. Despite cohort heterogeneity, the dysbiosis of lesion-adjacent tissues might provide us with further perspectives in demonstrating the role of the microbiota in colorectal cancer tumorigenesis.

IMPORTANCE Turbulent fecal and tissue microbiome dysbiosis of colorectal carcinoma and adenoma has been identified, and some taxa have been proven to be carcinogenic. However, the microbiomes of surrounding adjacent tissues of colonic cancerous tissues were seldom investigated uniformly on a large scale. Here, we characterize the microbiome signatures and dysbiosis of various colonic cancer sample groups. We found a high correlation between colorectal carcinoma adjacent tissue microbiomes and their on-site counterparts. We also discovered that the microbiome dysbiosis in adjacent tissues could discriminate colorectal carcinomas from healthy controls effectively. These results extend our knowledge on the microbial profile of colorectal cancer tissues and highlight microbiota dysbiosis in the surrounding tissues. They also suggest that microbial feature variations of cancerous lesion-adjacent tissues might help to reveal the microbial etiology of colonic cancer and could ultimately be applied for diagnostic and screening purposes.

KEYWORDS 16S rRNA, colorectal cancer, adenoma, carcinoma, adjacent tissues, network, supervised learning, enterotype, tissue microbiome, microbial ecology, random forest

Citation Mo Z, Huang P, Yang C, Xiao S, Zhang G, Ling F, Li L. 2020. Meta-analysis of 16S rRNA microbial data identified distinctive and predictive microbiota dysbiosis in colorectal carcinoma adjacent tissue. mSystems 5: e00138-20. https://doi.org/10.1128/mSystems.00138-20.

Editor Vanni Bucci, University of Massachusetts Medical School

Address correspondence to Fei Ling, fling@scut.edu.cn, or Lin Li, lilin@genomics.cn.

Received 20 February 2020
Accepted 25 March 2020
Published 14 April 2020

Colorectal cancer was the third most diagnosed cancer (10.2% of total cases) and the second leading cause of cancer death (9.2% of total cases) worldwide for both sexes in 2018 (1). In 2019, colorectal carcinoma (CRC) was estimated to be the third leading cancer type for both new cases and deaths in the United States, with over 140,000 diagnosed and about 50,000 dead combined in men and women (2). Accurate and early diagnosis is crucial in cancer treatment. Apart from the fecal immunochemical test (FIT), the guaiac-based fecal occult blood test (g-FOBT), and colonoscopy, attempts to investigate the stool microbiota to detect colonic carcinomas and adenomas have been proven to be potentially feasible (3). However, the complete process of how the microbiome interacts with colorectal cancer adjacent tissues is not fully understood.

The development of amplicon and shotgun genome sequencing technologies enabled a better understanding of the relationship between the microbiota and the host. For CRC and colorectal adenoma (CRA), many genera and species have been found to be significantly and consistently enriched or depleted in fecal samples compared with healthy individuals (3, 4). Due to the amplification procedure for DNA preparation, the 16S rRNA protocol outperformed the shotgun method in revealing the tissue microbiome composition (5). By comparing tumor tissue (on-site tissue) samples against the surrounding adjacent tissue samples (off-site tissues), differentially abundant microbial biomarkers were identified (6–8). Moreover, attempts at using the fecal microbiome to detect CRC noninvasively have been put into investigations (5, 9, 10).

However, there may be deficiencies in discriminating disease from the control using the fecal microbiota only. Several meta-analyses have been performed to study microbiome consistency and accuracy based on data sets derived from both amplicon and shotgun genome sequencing (3, 4, 6, 11, 12). Nonetheless, the classification of patients and healthy individuals using fecal microbiota methods was affected by confounders such as ethnic group, diet, and germ line genetic differences of individuals and inherent differences between fecal and tissue microbiomes. Thus, for the fecal microbiota, there is still a long way to transcend the existing screening method (13, 14).

In fact, the tissue microbiome was found to be different from the fecal counterpart (6, 14, 15). Usually, the mucosa or tissue might serve as the perfect environment for specific microbiota to come into effect during tumorigenesis (16, 17). This is especially the case for precancerous lesions, whose fecal microbial dysbiosis is moderate (10). Although several studies had revealed numerous disease-specific species and genera in on-site tissues (6, 18, 19), they have only seldom focused on disease-adjacent ones, which have been found to be difficult to be distinguished from their on-site counterparts using supervised learning methods (6). Moreover, the amplicon sequencing and analysis procedures differ from study to study regarding the hypervariable region, sequencing platform, sequence depth, and bioinformatics pipeline, making it even more challenging to be analyzed uniformly and systemically. Additionally, the results from a previous meta-analysis also showed that fine-scale classification of reads into operational taxonomic units (OTUs) did not help to improve the supervised learning classification performance significantly (12).

In this study, we chose to classify and assign taxonomy annotations to each filtered 16S rRNA gene read using the Kraken2 algorithm (20), making each level of taxa a feature and obtaining feature relative abundances for each sample independently. As a start, we performed a meta-analysis of 15 cohorts of colorectal carcinoma and adenoma samples. After consistent data preprocessing, we obtained feature relative abundances for downstream analyses. First, with the batch effect adjusted, we identified significantly abundant features in different case-control comparisons. Second, after pooling data from different cohorts, we trained random forest (RF) models and evaluated the performance of models in discriminating different sample groups. Importantly, we characterized the pattern of feature relative abundance alterations among sample groups. Third, we computed the correlation coefficient matrices that represented the ecology network and compared the network similarities among groups using the Mantel test. Finally, we investigated cohort heterogeneity and its impact on model classification using the Dirichlet multinomial mixture (DMM) method and cohort-to-

TABLE 1 Sizes of the large-scale 16S rRNA data sets included in this study[a]

Data set	No. of samples								Total	Sequencing platform	Sequencing region
	G-1	G-2	G-3	G-4	G-5	G-6	G-7	G-8			
Zeller	50	38	41				48	48	225	Illumina_MiSeq	V4
Flemer	62	23	69	59	31	2	74	65	385	Illumina_MiSeq	V3_V4
Burns							44	44	88	Illumina_MiSeq	V5_V6
Baxter	172	198	120						490	Illumina_MiSeq	V4
China_GBA				61	52	52	52	52	269	P_454	V1_V4
MAL2							21	23	44	Illumina_MiSeq	V3_V4
MAL1				6			20	21	47	Illumina_MiSeq	V3_V4
Zackular	30	30	30						90	Illumina_MiSeq	V4
Kostic							60	55	115	P_454	V3_V5
PNAS				11	2	2	23	23	61	P_454	V3_V5
Brazil				18				18	36	ION_TORRENT	V4_V5
China_KM							8	8	16	P_454	V1_V2
China_SH				20	31	31			82	Illumina_MiSeq	V3_V4
China_QD	11		10						21	Illumina_HiSeq_2500	V3_V4
China_SHTJ							65	65	130	Illumina_MiSeq	V4
Total	325	289	270	175	116	87	415	422	2,099	NA	NA

[a]G-1, Normal_Stools (healthy control stool samples); G-2, CRA_Stools (colorectal adenoma stool samples); G-3, CRC_Stools (colorectal carcinoma stool samples); G-4, Normal_Tissue (healthy control colorectal tissue); G-5, CRA_Tissue_Adjacent (colorectal adenoma adjacent tissue); G-6, CRA_Tissue (colorectal adenoma tissue); G-7, CRC_Tissue_Adjacent (colorectal carcinoma adjacent tissue); G-8, CRC_Tissue (colorectal carcinoma tissue); NA, not available.

cohort (C2C) and leave-one-cohort-out (LOCO) random forest models. Here, we identified distinctive and predictive microbial dysbiosis in the surrounding tissues of on-site colorectal cancer tissues. Importantly, the high similarity between the on-site and off-site colorectal cancer tissue microbiome signatures might provide us with novel perspectives in investigating the tumorigenic role of the microbiota along with the development of colorectal cancer disease.

RESULTS

Grouping of colorectal cancer microbiota data sets. Fifteen 16S rRNA data sets were retrieved from publicly available publications. Patients with CRC and CRA and healthy controls were included in this study, with CRC and CRA collectively called lesions in the following demonstration for convenience. Fecal and on-site and off-site lesion tissue samples were included. Detailed information on the data sets regarding sample size and others is depicted in Table 1.

Principal-coordinate analysis shows a distinct pattern of clusters. In our ordination analysis based on Bray-Curtis dissimilarity, extensive variations concerning sample groups were observed (Fig. 1a). For the tissue microbiota, principal-coordinate analysis (PCoA) showed distinguishing distributions between lesion, lesion-adjacent tissue, and normal control groups (Fig. 1b; see also Fig. S2B and C in the supplemental material), while for the adenoma stool group, the distribution was not significantly different from that of the normal stool group ($P = 0.211$ for Adenoma_Stools-VS-Normal_Stools as determined by analysis of molecular variance [AMOVA]) (Fig. S2D). Visible separation among cancerous lesion tissues, lesion-adjacent tissues, and healthy colon tissues indicated that the underlying ecological discrepancy among them could be distinguishable.

Differential abundance analysis identifies significantly enriched and depleted features in various case-control strategies. We deployed eight differential abundance analysis (DAA) strategies, S1 (CRA_Stools-VS-Normal_Stool) ($n = 614$), S2 (CRA_Tissue-VS-CRA_Tissue_Adjacent) ($n = 203$), S3 (CRA_Tissue-VS-Normal_Tissue) ($n = 262$), S4 (CRA_Tissue_Adjacent-VS-Normal_Tissue) ($n = 291$), S5 (CRC_Stool-VS-Normal_Stool) ($n = 595$), S6 (CRC_Tissue-VS-CRC_Tissue_Adjacent) ($n = 837$), S7 (CRC_Tissue-VS-Normal_Tissue) ($n = 597$), and S8 (CRC_Tissue_Adjacent-VS-Normal_Tissue) ($n = 590$) (where VS stands for "versus"), to investigate the potential microbiota differences. In our CRA_Stools-VS-Normal_Stools strategy, only a few features were found to be significantly enriched or depleted (Fig. 2a). Consistent with data from previous studies,

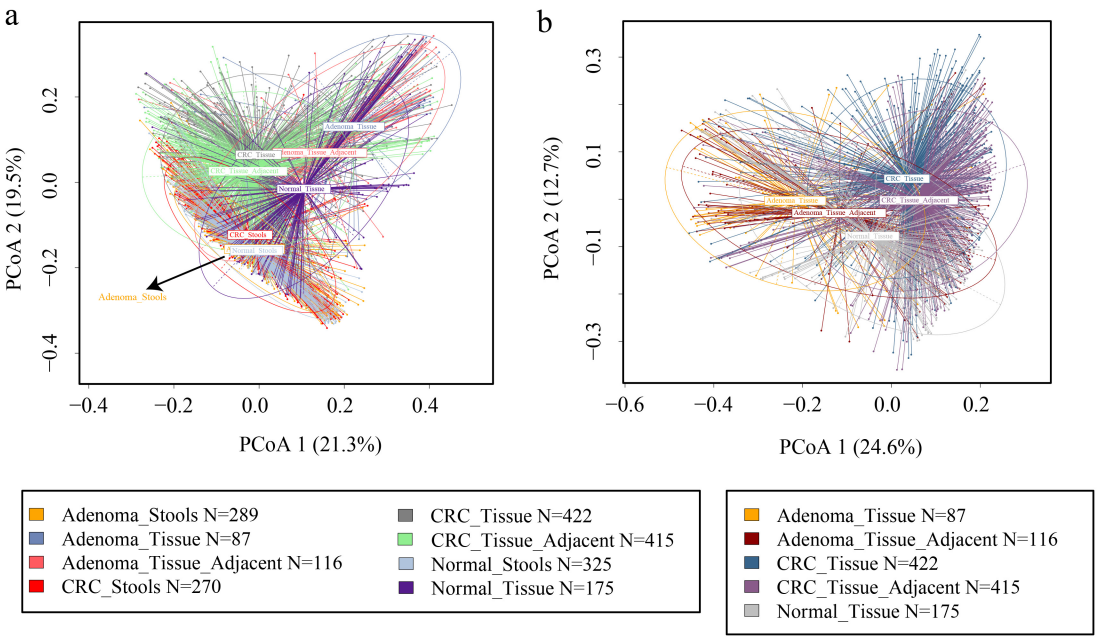

FIG 1 Principal-coordinate analysis (PCoA) in viewing groups of samples. (a) The beta diversity based on the Bray-Curtis metric was used to perform PCoAs. The first two principal coordinates were graphed to visualize the sample group relationships. (b) A similar procedure was applied to CRA and CRC tissue-associated samples to demonstrate particular sample relationships. Each ellipse in different colors represents 95% of the inertia of the corresponding group. Group tags that are not legible due to overlap are specified using dark arrows.

compared to healthy stool samples, *Porphyromonas endodontalis* (false discovery rate [FDR]-adjusted *P* value of 1.28e−20), *Fusobacterium* (FDR-adjusted *P* value of 1.07e−36), *Prevotella intermedia* (FDR-adjusted *P* value of 9.23e−7), and *Parvimonas* (FDR-adjusted *P* value of 7.87e−65) were found to be significantly enriched in CRC stool samples (Table S2) (4, 12), while only 9 features were founded to be significantly depleted (Fig. 2a). There were 63 and 142 enriched and 42 and 41 depleted features in the CRA_Tissue and CRC_Tissue groups compared to the normal tissues. However, 26 enriched and 41 depleted features could also be observed by strategy S4, while for S8,

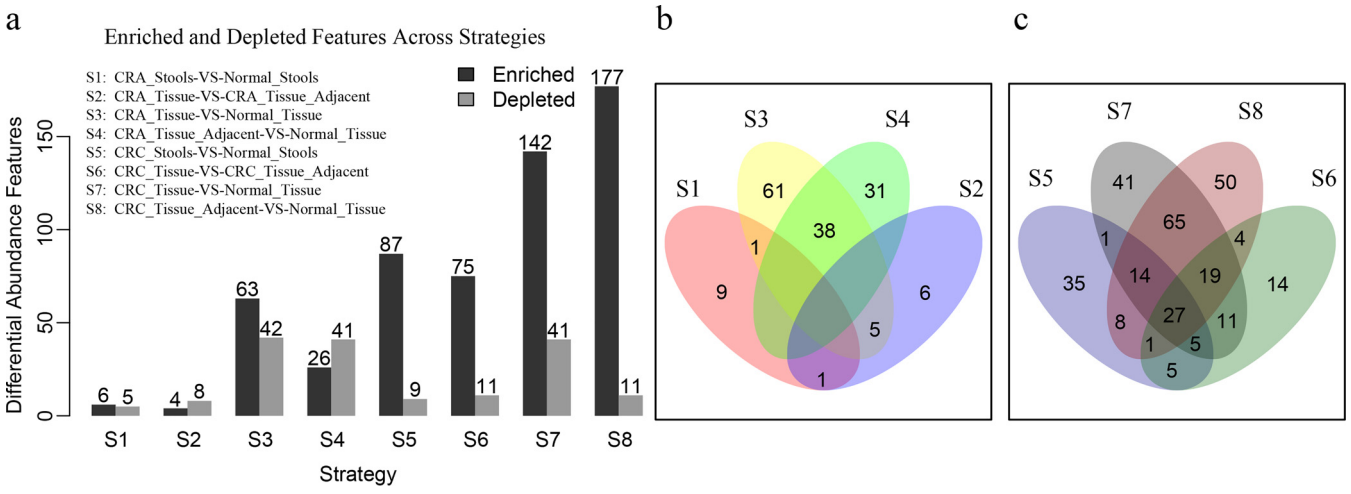

FIG 2 Differential abundance analyses identify enriched and depleted features in different strategies. (a) Differential abundance analysis was conducted by using the DEseq2 software algorithm for each strategy, with an absolute $\log_2$-fold change above 1 and an adjusted *P* value of less than 0.05 considered enriched or depleted. The cohort was added as a factor to adjust the batch effect. The operator "-VS-" was organized in a form such that the former was compared against the latter, with the enrichment and depletion concepts based on the former. (b and c) Venn plots showing the overlap of the features among different strategies for adenoma- and carcinoma-associated diseases.

the numbers were 177 and 11, respectively. This led us to find 38 and 125 overlapping features between strategies S3 and S4 and between S7 and S8, respectively (Fig. 2b and c). For the consideration that a large number of overlaps might simply be due to the fact that the adjacent cancer tissues and cancer tissues were paired and derived from the same individual, we conducted differential abundance analysis by keeping 336 pairs of carcinoma patients and 83 pairs of adenoma patients separated (Fig. S3A). We still found many overlapping features between strategies S3 and S4 and between S7 and S8 (Fig. S3B and C). All results suggested that the microbiota shared between lesions and their adjacent tissues could be undervalued. Detailed results for DAA features in each strategy are shown in Table S2.

Microbiome CRC and CRA classification models. To learn about the extent to which the microbial components were different among sample groups and the capacity of the use of microbial information to discriminate colorectal neoplastic diseases, we established random forest (RF) classifiers for all eight strategies by pooling samples (pooling RF model). We decoded the sequencing platform and 16S rRNA hypervariable region information as binary features and estimated their effects on our RF models. Because other factors such as colon preparation method, sample collection, and DNA preparation before sequencing were highly heterogeneous and difficult to standardize, we included cohort factor as a binary feature to assess the impact of cohort heterogeneity. Features maximizing the AUC (area under the curve) value or making the AUC value reach a plateau were selected (Fig. S4A). Additionally, we conducted the same pooling RF analysis without adding these binary features and achieved a similar performance (Fig. S3B). The importance of binary features was relatively low, except for strategy S3 (Fig. S3C).

When using all the CRC and control stool samples to train and predict CRC with 10-fold cross-validation, the AUC was 80.7% (95% confidence interval [CI], 77.2% to 84.2%) in our model containing 61 features. The performance of our model was similar to that in recently reported research based on shotgun sequencing data measured by the AUC (3, 4, 11). Interestingly, when training and predicting CRA using stool samples, the best AUC was 69.9% (95% CI, 65.8% to 74.0%), which showed no deficiency compared with previous studies using shotgun sequencing (Fig. 3a) (3, 4).

In our CRA_Tissue-VS-Control_Tissue and CRC_Tissue-VS-Control_Tissue random forest models, the AUC values were 91.5% (95% CI, 88.3% to 94.8%) and 96.0% (95% CI, 94.7% to 97.4%), respectively (Fig. 3b and c). The top-ranked features of the former model were *Enhydrobacter*, *Streptococcaceae*, *Pseudomonas stutzeri*, *Psychrobacter*, *Streptococcus*, *Peptococcaceae*, and *Gluconacetobacter*, while the features *Blautia obeum*, *Dorea*, *Ruminococcus*, *Lachnospiraceae*, *Blautia*, mitochondria, *Parvimonas*, and *Fusobacterium* showed high importance scores in the latter model.

The CRA_Tissue_Adjacent-VS-Normal_Tissue model gave an outstanding performance, with an AUC value of 89.5% (95% CI, 86.0% to 93.0%) (Fig. 3b), and a better performance can be observed for the CRC_Tissue_Adjacent-VS-Normal_Tissue model (AUC = 95.8%; 95% CI, 94.4% to 97.2%) (Fig. 3c). In the CRC_Tissue-VS-CRC_Tissue_Adjacent and CRA_Tissue-VS-CRA_Tissue_Adjacent models, the AUC values were 77.0% (95% CI, 70.5% to 83.4%) and 74.8% (95% CI, 71.6% to 78.1%), respectively. A similar trend was reported in a previous study (6). Selected features for each strategy are shown in Table S3. There was no significant difference between the Normal_Tissue-VS-CRC_Tissue and CRC_Tissue_Adjacent-VS-Normal_Tissue strategy receiver operating characteristic (ROC) curves ($P = 0.824$ by DeLong's test) for either adenoma phenotype ($P = 0.402$ by DeLong's test). However, the use of the adjacent tissue microbiome was significantly better than the use of the stool microbiome for both carcinoma ($P = 2.797e{-}14$ between strategies S5 and S8 by DeLong's test) and adenoma ($P = 1.955e{-}12$ between strategies S1 and S4 by DeLong's test) detection.

Microbiota cross talk is widespread between models based on lesion tissues and lesion-adjacent tissues. Although both carcinoma and adenoma diseases could be efficaciously predicted or discriminated from healthy controls using stool, on-site

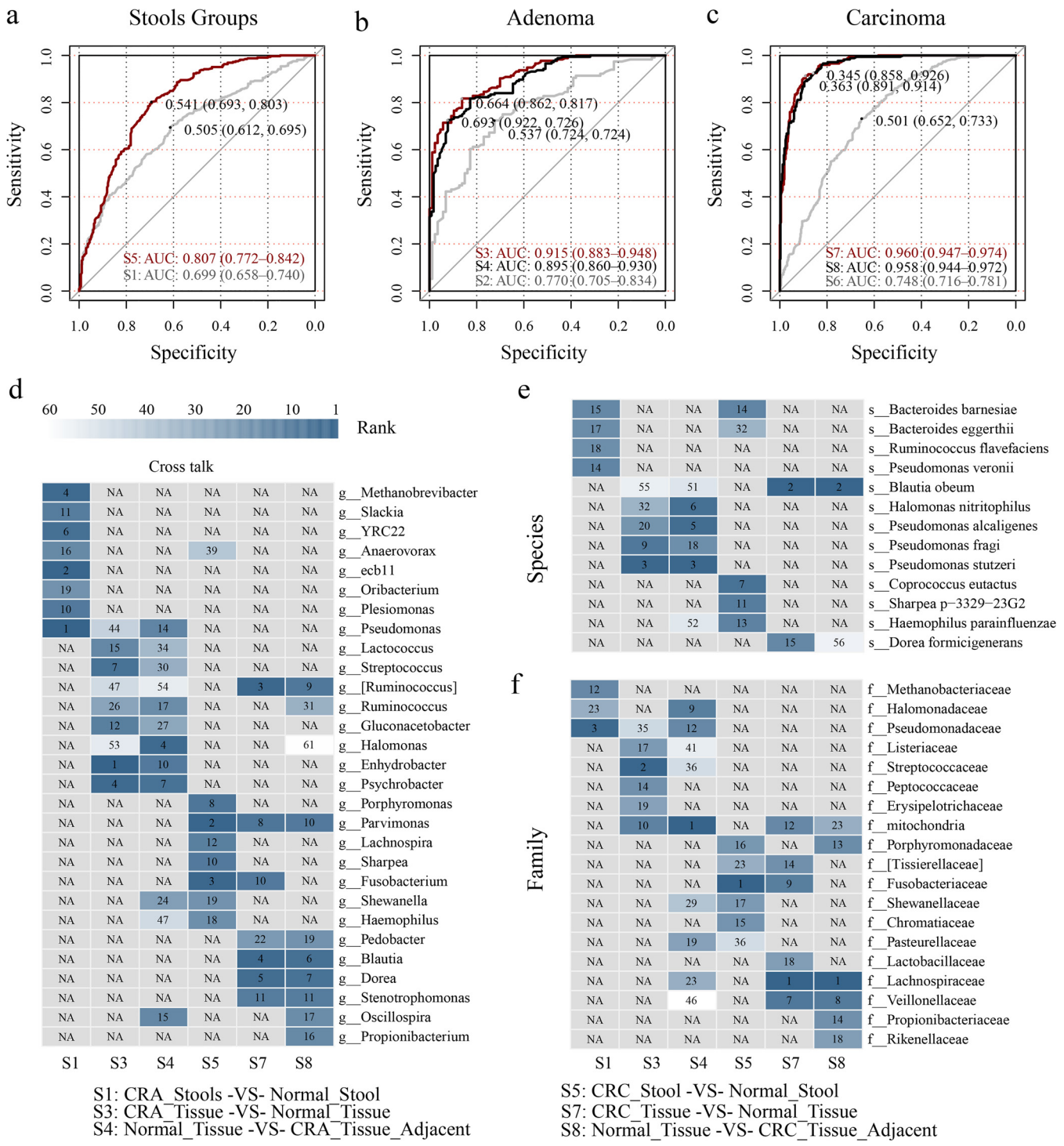

FIG 3 Pooling random forest models and feature cross talk. (a) Random forest models in predicting adenoma and carcinoma using stool samples with 10-fold cross-validation. (b and c) Pooling random forest models in predicting adenoma and carcinoma using tissue samples. (d to f) Genus-, species-, and family-level features' ranks were obtained according to the random forest model's MDIA and plotted in a heat map, with dark steel blue for the highest rank, white for the lowest rank, and gray for those that were unavailable. All the strategy codes were carried on as described above. All specificity and sensitivity thresholds were decided by the Youden index. NA, not available.

tissue, and lesion-adjacent tissue samples in our pooling random forest models to different degrees, we found that some features consistently ranked highly in all or most of the strategies in discriminating specific lesions. For instance, *Parvimonas* ranked highly in all three strategies in predicting CRC, while *Ruminococcus*, *Stenotrophomonas*,

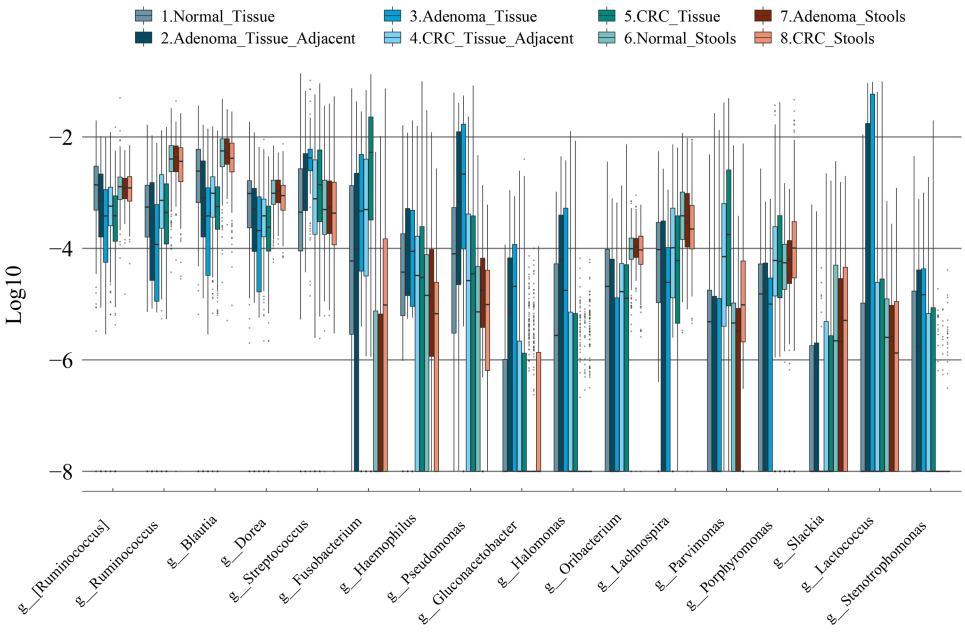

FIG 4 Consistent and divergent genus-level feature profiles along with the temporal order. Box plots of highly cross-talking genus-level features showing consistent and divergent profiles across different groups were arranged in theoretical disease development time series for both tissue and stool samples. A relative abundance of zero was set to 10e−9 to avoid an infinite number.

Blautia, *Dorea*, and some other genera were shared by strategies S7 and S8, which used carcinoma on-site tissues and the surrounding adjacent tissues. As for the prediction of adenoma, *Enhydrobacter*, *Psychrobacter*, *Lactococcus*, and *Pseudomonas* were those that cross talked actively between strategies S3 (CRA_Tissue-VS-Normal_Tissue) and S4 (CRA_Tissue_Adjacent-VS-Normal_Tissue) (Fig. 3d). Additionally, some species-level and family-level features also harbored high importance as measured by the mean decrease in accuracy (MDIA) rank (Fig. 3e and f).

Parallel and divergent feature enrichment and depletion patterns might help reveal microbial distribution across disease development. Adenoma, the precursor of the majority of CRCs (21), was less malignant than CRC and harbored microbiome profiles different from those of CRC (10). We arranged our feature relative abundance profiles in the assumptive Normal_Tissue, CRA_Tissue_Adjacent, CRA_Tissue, CRC_Tissue_Adjacent, and CRC_Tissue temporal order for our presentation. In our pooling analyses, microbial feature relative abundance evidence supported the continuous alteration patterns of the microbiota in CRAs and CRCs. For instance, some taxa, like *Fusobacterium*, showed a gradual accumulation, while others, like *Ruminococcus*, *Blautia*, and *Dorea*, showed progressive depletions along the temporal sequence (Fig. 4). However, we also found some features that were enriched in adenoma tissues and the corresponding adjacent tissues but not prevalent in carcinoma-associated tissues, such as *Streptococcus*, *Haemophilus*, *Pseudomonas*, *Gluconacetobacter*, *Halomonas*, and *Lactococcus*. Similarly, compared with healthy tissues, *Parvimonas* and *Porphyromonas* were enriched in CRC tissues and CRC adjacent tissues but not dominant or even depleted in adenoma-related tissue groups (Fig. 4). These results indicated that the dynamic community change might reflect the succession of different microbiota during colonic cancer development.

Coabundance analysis identifies highly correlated microbial networks between lesions and corresponding adjacent tissues. We inferred the taxon-taxon correlation coefficient matrix for each type of sample with all filtered features using SparCC (22) software. After hierarchical clustering, features harboring similar coefficient profiles gathered together as clusters, which is referred to as the coabundance group here (Fig. 5a) (8). First, adenoma adjacent tissues harbored the largest number of positive

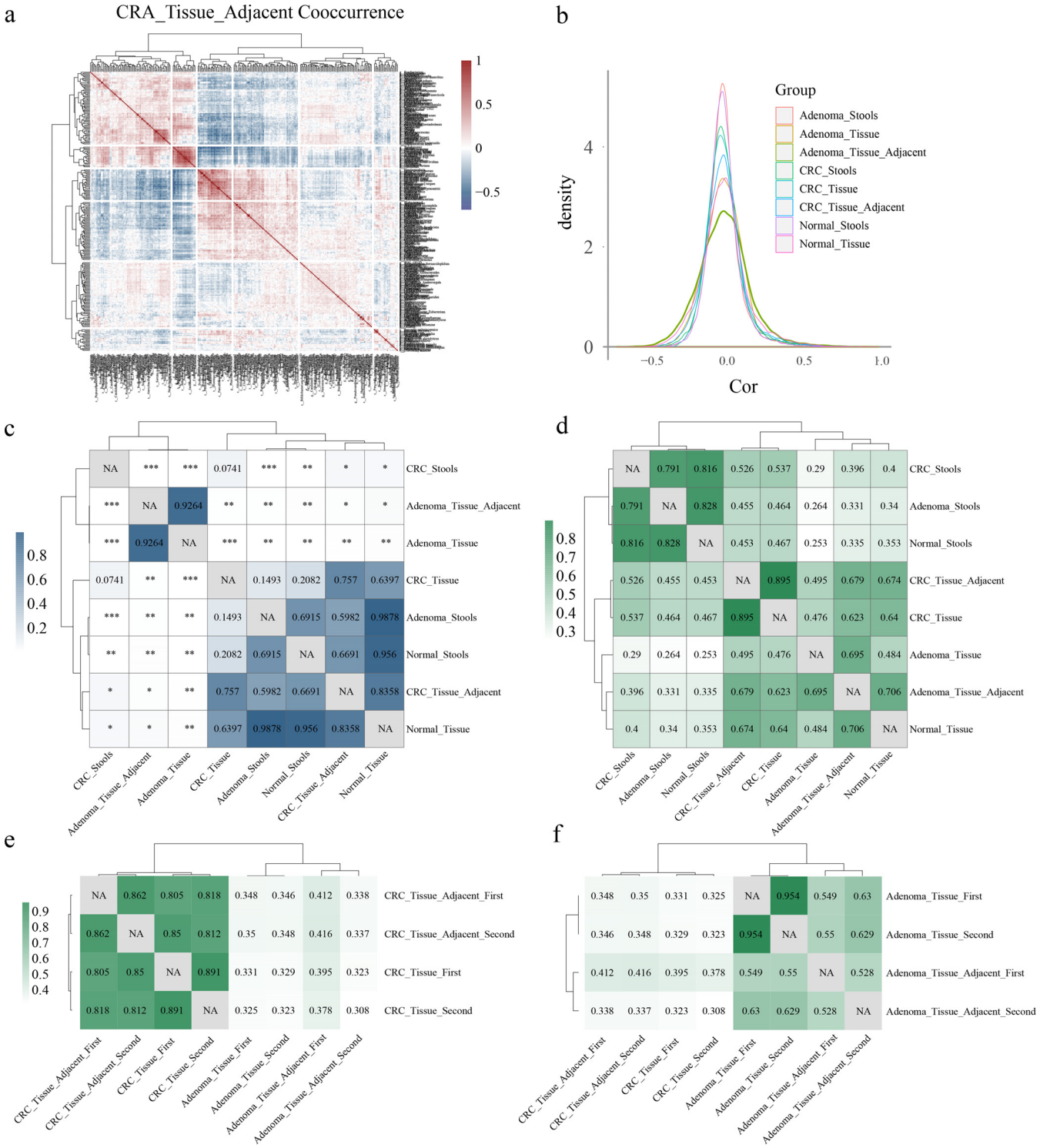

FIG 5 Taxon-taxon correlation coefficient profiles. (a) The coabundance group in adenoma adjacent tissue samples based on the SparCC correlation coefficient was hierarchically clustered to show cooccurrence features. (b) Kernel density estimates of correlation coefficients derived from SparCC for the eight strategies. (c) *P* values of correlation intensions were calculated between different types of samples using correlation coefficient intensity data by the Wilcoxon rank sum test. ***, $P < 0.001$; **, $0.001 < P < 0.01$; *, $0.01 < P < 0.05$. (d) Correlation of networks among different samples. The Mantel test was performed between different types of samples using the correlation coefficient matrix to identify the Mantel *r* statistic representing the extent to which the two matrices were correlated. All the correlations presented were significant, and *P* values are not shown ($P < 0.001$). (e and f) The Mantel test was performed with paired lesion tissues and lesion-adjacent tissues separated.

and negative correlations above the absolute threshold of 0.3 (Fig. 5a and b). Second, after computing the kernel density estimation of the correlation coefficient intensity for each type of sample, we found that the higher coefficients were more prevalent in tissues than in stool samples (Fig. 5b). Consistent with the results of a previous study (23), we observed more correlations higher than 0.3 ($n = 975$) than those lower than -0.3 ($n = 191$) in the CRC_Stools groups. The same result could also be seen when all eight sample groups were aggregated together. To determine whether the differences in interaction strength (absolute value of the correlation coefficient) among sample groups were significant, *P* values were calculated using the Wilcoxon rank sum test. Neither the overall interaction strength between CRA_Tissue_Adjacent and CRA_Tissue ($P = 0.926$) nor that between CRC_Tissue_Adjacent and CRC_Tissue ($P = 0.757$) was significantly different (Fig. 5c).

Using the Mantel test, correlations between each of two networks were calculated. The stool samples showed high correlations, with correlation coefficients between CRC_Stools and Normal_Stools, Adenoma_Stools and Normal_Stools, and CRC_Stools and Adenoma_Stools being 0.816, 0.828, and 0.791 (all $P < 0.001$), respectively. Surprisingly, the correlations between CRC_Tissue and Normal_Tissue, Adenoma_Tissue and Normal_Tissue, and CRC_Tissue and Adenoma_Tissue were 0.64, 0.484, and 0.476 (all $P < 0.001$), respectively (Fig. 5d). This result suggested that microbial networks were more divergent in tissue groups than in stool groups, which was consistent with our random forest discriminating results (Fig. 3a to c). Among the correlation values between carcinoma tissues and other groups, the highest one was 0.895 (between CRC_Tissue and CRC_Tissue_Adjacent), while for the adenoma tissues, the highest value was 0.695 (between Adenoma_Tissue and Adenoma_Tissue_Adjacent). After separating paired samples from the same individuals, as depicted in Fig. S3A, we still observed high correlations between the CRC_Tissue and CRC_Tissue_Adjacent and between the Adenoma_Tissue and Adenoma_Tissue_Adjacent groups (Fig. 5e and f).

Metacommunity partition reveals cohort-specific patterns of microbial ecology in tissues. When using the DMM method to decide the optimized number of metacommunities or partitions in tissue samples, including normal controls, carcinoma tissues, adenoma tissues, and their corresponding surrounding adjacent ones, the best partition number was 9 under the Bayesian information criterion (BIC) estimate (Fig. S5A). The sample number and percent distribution in each metacommunity are depicted in Fig. S5B and C.

Harboring the lowest percentage ($n = 11$; 7.6%) of carcinoma on-site tissues, metacommunity A mainly contained benign samples ($n = 133$; 92.4%) and showed low chao1 alpha diversity. Features such as *Fusobacterium* (24) and *Parvimonas*, which were thought to be highly correlated with carcinogenesis, showed low relative abundances (Fig. S5F). Containing the largest number of carcinoma adjacent samples ($n = 91$; 52.9%), metacommunity B showed a sparse microbial ecology, represented by low chao1 alpha diversity as well (Fig. S5D). Consisting of the second-highest percentage of benign samples ($n = 74$; 91.4%), metacommunity C contained no normal tissue and harbored the highest relative abundances of *Lactococcus* and *Pseudomonas*. Metacommunity D, in which 88.5% of the samples were carcinoma related, showed the highest average chao1 (Fig. S5D) and relatively high Simpson alpha diversity (Fig. S5E) values. In metacommunity E, 23.7% ($n = 40$) of the samples were from the normal tissue group, with the rest of the samples being carcinoma related. In metacommunity F, which showed the highest average Simpson alpha diversity value, 91.1% of samples were carcinoma related.

Similar to metacommunity C, metacommunity F showed high abundances of *Lactococcus* and *Pseudomonas*. Both metacommunities C and F were mainly from China (China_SH [China, Shanghai] [38] and China_SHTJ [China, Shanghai Tongji] [40]) and prevalently harbored adenoma-related and carcinoma-related samples, respectively. In metacommunity G, we found the largest number of normal control tissue samples ($n = 53$). Enrichments of microbiota taxa like *Dorea*, *Ruminococcus*, and *Blautia* and depletions of *Fusobacterium* and *Parvimonas* (Fig. S5F) were also identified. Collectively,

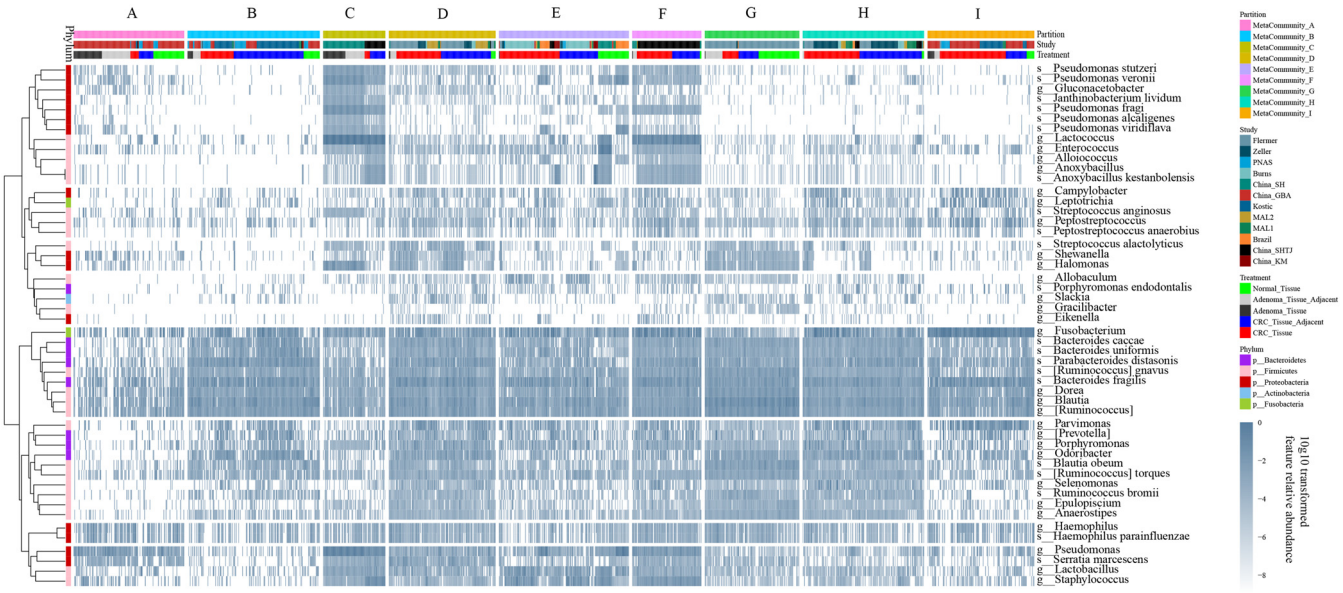

FIG 6 The DMM method identifies different enterotypes representing the microbiome profile. All 657 filtered features of 1,215 tissue samples, including normal samples, adenomas and carcinomas, and the corresponding adjacent tissues, were subjected to the DMM model, resulting in 9 metacommunities. Genus- and species-level taxa that overlapped in DEseq2 and pooling random forest models across 8 strategies were hierarchically clustered using the Pearson correlation and are presented in rows. The feature relative abundance was $\log_{10}$ transformed, with an abundance of zero set to 10e−9 to avoid infinite numbers.

both metacommunities A and G harbored a large number of healthy tissue samples and were represented by low and high chao1 diversity values, respectively. Interestingly, features like *Ruminococcus*, *Blautia*, and *Dorea* showed low abundances in the former but high abundances in the latter. Like metacommunity D, 96.8% of samples ($n = 153$) in metacommunity H were carcinoma related. Both metacommunities D and H might represent a kind of widely existing metacommunity of high alpha diversity ecology in the colorectal cancer disease population (Fig. S5D and E). Metacommunity I, which mainly contained the CRC on-site tissue samples ($n = 86$; 61.9%), was characterized by the prevalence of *Fusobacterium* and *Parvimonas* (Fig. 6 and Fig. S5F) and showed substantially low chao1 metric alpha diversity (Fig. S5D).

The cohort-to-cohort random forest model achieves better internal cohort classification. Although cohort information has been taken into consideration in pooling RF models and DEseq2 analyses, cohort-specific metacommunities were dominant, as shown by DMM analysis, regardless of disease status (Fig. 6). In order to characterize the reproducibility of our conclusion drawn by pooling all samples, we conducted random forest analysis with 10-fold cross-validation in each cohort and used the others as independent validation data sets in each strategy, here called the cohort-to-cohort random forest model (C2C RF model). Models utilizing adjacent tissues could discriminate adenoma and carcinoma from healthy tissues with AUC values of 0.90 and 0.95 on average in the training module (Fig. 7a and b).

However, weakness appeared in cross-cohort validations. Similar trends of high AUC values in training data sets but inferior AUC values in validation data sets could be observed for other strategies (Fig. S6A to F). Furthermore, when leaving one cohort out as the independent validation data set and training the rest in the CRC_Tissue_Adjacent-VS-Normal_Tissue strategy, better validation performance could be seen when the MAL1 and PNAS cohorts were left out for validation (Fig. 7c to f). In the CRA_Tissue_Adjacent-VS-Normal_Tissue strategy, when leaving cohorts China_GBA and Flemer out as validation data sets, the training model achieved AUC values of 0.939 and 0.90. When leaving the China_SH cohort out as the independent validation, we observed AUC values of 0.829 and 0.978 in the training and validation models, respectively (Fig. 7g to i).

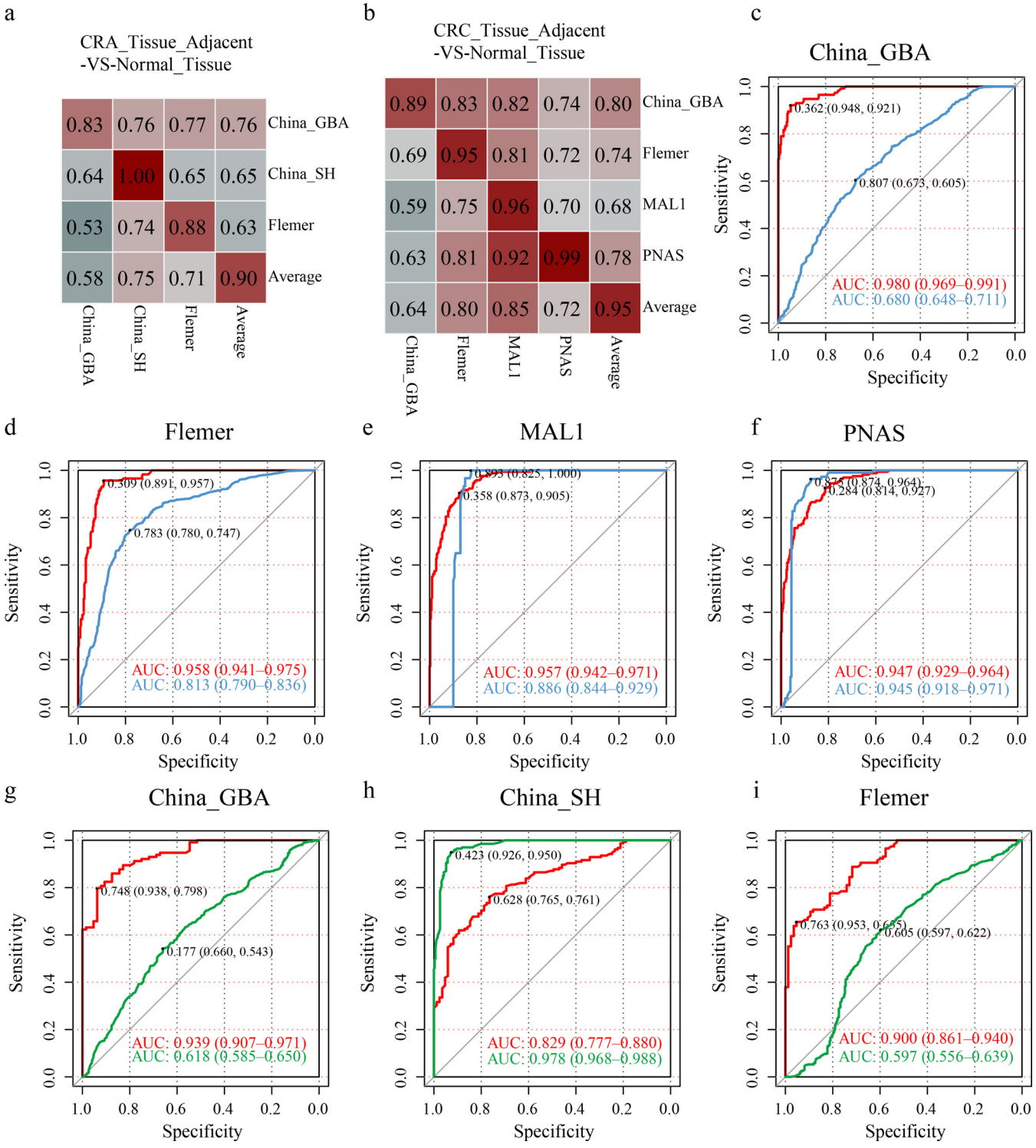

FIG 7 Cohort-to-cohort and leave-one-cohort-out random forest models. Classifier performances in the 10-fold cross-validation model within each cohort (along the diagonal, with the last number representing the average AUC) and the cohort-to-cohort training-testing models were measured by the AUC (off-diagonal, with the top *n*-1 row and column representing the training-testing data set, respectively [*n* is the cohort number]). The last row and column depict the average AUCs of the testing data sets to demonstrate the generalization ability to be predicted by multiple cohorts and predicting others. (a and b) Cohort-to-cohort performance for strategies CRA_Tissue_Adjacent-VS-Normal_Tissue and CRC_Tissue_Adjacent-VS-Normal_Tissue. (c to f) LOCO random forest models in training and testing carcinoma against normal samples using microbiota information of adjacent tissues. (g to i) LOCO random forest models in training and testing adenoma against normal samples using adjacent tissue microbiota information.

DISCUSSION

Here, we confirmed that lesion-adjacent tissues were not as healthy as normal tissues, which was mentioned in previous small-sample-size research (7). Through PCoA, we identified distinct distributions of adjacent tissues compared with other sample groups. Second, we found a large number of overlapping features in strategies that discriminated lesion tissues and the lesion-adjacent tissues from normal tissues. Next, we found that the microbiome of lesion-adjacent tissues played an important role in supervised machine learning models that discriminated lesions from controls. We validated our hypothesis by calculating the network correlations between different sample groups, especially between lesion tissues and the surrounding adjacent ones. Finally, we showed that despite cohort heterogeneity, the microbial dysbiosis of lesion-adjacent tissues was validated to be a widespread phenomenon.

Compared to a single study, pooling of data sets from multiple studies enabled us to detect comprehensive alterations by strengthening the signal of relative abundance and reducing false positives with a comparable strict filtering standard to reject low-frequency taxa. When pooling samples from different cohorts, compared with the shotgun sequencing method using stool samples, both CRC and CRA models based on the 16S rRNA data set showed no distinct deficiency in the prediction (25). We characterized that lesion-adjacent and on-site lesion tissues could not be efficiently discriminated, as previously reported (6). As illustrated in our network analysis, the high microbial network correlation between lesion and surrounding lesion-adjacent tissues indicated that the microbial network configurations between them were highly similar (Fig. 5d to f). Besides network differences, a dynamic change of the relative abundance of a single feature in temporal order might help to identify driving tumorigenic factors in CRC development. Low relative abundances of *Ruminococcus*, *Blautia*, and *Dorea* were also reported in cancerous tissues (15). Interestingly, the dynamic fluctuation of the relative abundances of *Pseudomonas*, *Streptococcus*, *Porphyromonas*, and *Fusobacterium* in these sample groups, especially in adjacent tissues, might pave the way toward understanding their roles in tumorigenicity (Fig. 4).

Cohort heterogeneity is critical in affecting transcohort generalization. In pooling RF models, when some factors were included as binary features, we found that the China_GBA cohort factor ranked high in strategy S3: Adenoma_Tissue-VS-Normal_Tissue. This might be because 60% (52 out of 87) of adenoma tissue samples were from cohort China_GBA, whose data were from the 454 sequencing platform using the V1-V4 regions (see Fig. S4C in the supplemental material). In strategy S8, although the cohort China_GBA was inferior in discriminating CRC_Tissue_Adjacent from Normal_Tissue compared to cohorts MAL1, PNAS, and Flemer in the training model, it had a higher AUC score (0.80 on average) in independent testing cohorts. This suggested better training and testing generalizations (Fig. 7a and b). Particularly, when cohort China_GBA was left out as a validation data set, the poor performance in LOCO analysis in discriminating adenoma and carcinoma further confirmed cohort heterogeneity problems. Although limitations in machine learning performance were inevitable in the existing pipeline, especially in the adenoma stool-versus-control stool model (Fig. S6A to F), we still found comprehensive, reproducible features across cohorts in each strategy (Fig. S6G). For reproducible features, most of them were also identified in pooling random forest modules (Fig. S6G). All reproducible taxa in the eight strategies are summarized in Table S4.

Cohort heterogeneity was also observed regardless of disease status. High abundances of *Lactococcus* were observed in both the China_SH and China_SHTJ cohorts, suggesting that geography and ethnic groups were essential in shaping the specific microbiota community, as previously revealed (17), regardless of colorectal cancer disease status. We also observed that not each carcinoma sample was turbulent enough to be grouped into malignant groups, while some healthy samples were grouped with lesions. Some cohort-specific signatures harbored by metacommunity C might help to explain the high transcohort generalization testing results (Fig. 6h).

Future research including a large number of samples of a specific ethnic cohort is encouraged to characterize cohort-specific tissue microbiome signatures and explain the driving factors shaping them.

Interestingly, since the microbiome component was distinctive in adjacent tissues, it might serve as an alternative for colorectal cancer screening, specifically for sigmoid cancer. The excellent performance in predicting cancer using the microbiota of surrounding adjacent cancerous tissues showed its potential for clinical application. It is challenging to obtain colorectal tissues of screening participants. However, according to previous studies sampling the mucosal microbiome (26–28), colonic lavage fluid, aspirated luminal contents, or the loose mucus layer could serve as a relatively accurate proxy in providing biopsy specimen microbiota compositions. Here, we examined this possibility by illustrating the following problems. First, there was no significant difference between strategies S7 and S8, indicating that the use of the adjacent tissue microbiome was sufficient for disease predictions. Second, Youden's index maximizing the sum of sensitivities and specificities was applied to decide a threshold for the CRC_Tissue_Adjacent-VS-Normal_Tissue pooling RF model and achieved a sensitivity of 0.926 under a specificity of 0.858. For the CRA_Tissue_Adjacent-VS-Normal_Tissue model, the sensitivity was 0.726 under a specificity of 0.922 (Fig. 3b and c).

In the future application of the use of the gut microbiome to predict CRC/CRA, the mucosal microbiota might be an alternative and capable candidate for clinical application. Since some tissue samples used in our meta-analysis were not exactly mucosa but were biopsy specimens, lessons learned from currently available information remained incomplete. Further studies revealing the extent to which the distal mucosa microbiome represents the corresponding cancer-associated adjacent tissues are still needed. More investigations of the mucosa, especially the distal rectal mucosa microbiota, might help to develop a protocol that guides the sampling of the distal gastrointestinal tract mucosa noninvasively. For instance, mucosal-luminal interface (MLI) mucus (28), a mixture of the loose mucus layer sampled from the intestinal wall by washing off and aspirating, proved to harbor a biomass highly similar to that of biopsy specimens (29, 30) and might serve as a replacement for biopsy specimens. Other methods like colon swap could be used as an in-house device if qualified to capture crucial mucosal microbiome signatures. Furthermore, a new tool designed to scrape the colon mucosa by clinician rectum examination could also serve as an alternative instrument. In this way, carcinomas and adenomas located on the sigmoid colon might be detected with high sensitivities and specificities based on the knowledge that the lesion-adjacent mucosa microbiota plays a crucial role in making prediction models.

In our further research, some other improvements might help to obtain better performance. First, although the feature-based method could effectively utilize microbiome information in disease classifications, unlike *de novo* OTU picking protocols that took advantage of each filtered read, some reads were rejected by either Trimmomatic (31) or Kraken2 (20) in our pipeline, resulting in the omission of some unknown taxa which might be potential markers in specific case-control models. Second, the sample read numbers differed from cohort to cohort in magnitude, which made it difficult to perform rarefaction. To keep more details of read information, after checking that the feature number would reach a plateau in each cohort, read numbers were normalized to obtain relative feature abundances without rarefaction for downstream analyses. In the planned compatibility progress, the concordant sample preparation pipeline might help identify vital functional elements in different types of samples unfailingly.

Finally, we would like to integrate other possible confounding factors, like age, gender, body mass index (BMI), tumor location, methods for obtaining biopsy specimens, and cancer status, in our future studies. For unity consideration, mucosa and biopsy specimens were combined and termed "tissue" in our analyses. Moreover, we used a filtering pipeline to remove bacterial taxa existing in fewer than 10% of samples after read normalization. While minimizing the false-positive discovery rate, this might have led to the missing of some rare but interesting taxa.

In conclusion, we identified significant dysbiosis in both lesions and lesion-adjacent tissues compared with healthy colon tissues. We also found that judged from the microbiome component perspective, lesion-adjacent tissues should not be regarded as healthy colon tissues. This research provided new perspectives for further research in revealing the role of the microbiome in tumorigenesis along with the development of colorectal tumors.

MATERIALS AND METHODS

Data set collection and sequencing data preprocessing. The 16 data sets were labeled as Zeller (5), Flemer (32), Burns (33), Baxter (10), China_GBA (China Great Bay Area) (23), MAL2 (34), MAL1 (34), Zackular (9), Kostic (35), PNAS (36), Brazil (19), China_KM (China, Kunming) (37), China_SH (38), China_QD (China, Qingdao) (39), and China_SHTJ (40). Raw sequence data and metadata were retrieved from the NCBI or from the authors directly. All sequences were trimmed by using Trimmomatic (31) as described in previous research (41). For sequences generated by the Illumina platform, Illumina-specific adapters were removed using default parameters. Samples were excluded if metadata were not available. Samples were grouped before downstream analyses. First, stool samples derived from healthy controls and adenoma and CRC patients were grouped as Normal_Stools, CRA_Stools, and CRC_Stools, respectively. Second, tissue samples from healthy control and adenoma and carcinoma disease sites were grouped as Normal_Tissue, CRA_Tissue, and CRC_Tissue, respectively. Third, we grouped the surrounding tissue samples (usually 5 to 10 cm away from the lesion) that were adjacent to the cancerous sites as CRA_Tissue_Adjacent and CRC_Tissue_Adjacent for adenoma and carcinoma patients.

Sequence classification, taxonomy determination, and feature relative abundance. The Kraken2 (20) algorithm was applied to classify each high-throughput sequence read directly against the Green-Genes database and return their taxa. Each sample was processed independently to gain a mataphlan2 (42) format report of microbiome compositions, with features ranging from kingdom to species. The Kraken2 report was filtered under the criterion that each feature must exist in at least 10% of all samples.

Beta diversity and principal-coordinate analyses. Feature relative abundance-based information was subjected to calculation of the beta diversity using Bray-Curtis dissimilarity metrics via a module implemented in Qiime (43) software. Subsequently, we performed principal-coordinate analysis (PCoA) based on our Bray-Curtis dissimilarity matrix.

Use of the random forest machine learning model to discriminate sample groups. All the random forest models were built using the supervised_learning.py command in Qiime software (version 1.9.1) (43). This script was called by the randomForest R package (version 4.6-14) and was used to perform random forest analysis with default parameters using inner 10-fold cross-validation to avoid overfitting. All returned feature importance scores were characterized using MDIA to present the importance of features in model classifications. In the optimal feature number decision procedure, all features were included to obtain the importance of each feature, from which they were sorted. Next, the above-mentioned models were repeated, with previously ranked features added one by one, starting from the most important one. The optimized model that made the AUC value reach a plateau or peak was selected. Finally, all the resulting probabilities served as the input for the pROC R packages to compute the AUC values and draw the ROC (receiver operating characteristic). A similar feature selection procedure was applied to cohort-to-cohort and leave-one-cohort-out (LOCO) RF models, in which we added an additional prediction function to use independent data sets for validation to evaluate the generalization of trained models and gained each tested sample's probabilities of being assigned to different groups. In each circle of added features, the model that maximized the sum of training and validation AUCs was chosen, and the corresponding features were determined as potential markers for downstream analyses.

Feature cross talk. The importance of each feature was represented by ranking according to the MDIA in the pooling RF model using 10-fold cross-validation. Features ranking in the top 25 in the corresponding strategy were selected and separated into family, genus, and species taxonomy groups.

Correlation network inference. The filtered features (genus-level and species-level features that existed in at least 20% of 2,099 samples) were subjected to the SparCC (22) algorithm to calculate the taxon-taxon correlation coefficient matrix for each group of samples using default parameters. Correlation coefficient matrices of each group of samples were sorted in the same taxonomy order and were applied to compare network similarities.

Cohort-to-cohort reproducible features. For cohort-to-cohort RF models in each strategy ($n = 8$), each cohort served as a training data set and was tested using others, resulting in $n \times (n - 1)$ training-testing pairs, as demonstrated, with n representing the cohort numbers. For the reason that some training-testing pairs reached optimal status by maximizing the sum of the discovery AUC and the validation AUC using only a few features, features prevailing in at least 30% of pairs were regarded as highly reproducible.

Determination of enterotype for tissue samples. Relative abundance-transformed counts of 657 filtered features were subjected to the DMM algorithm (44) to identify groups of metacommunities harboring similar microbial configurations using Mothur (45) software with default parameters for tissue samples. Nine metacommunities were obtained based on the BIC approximation. Subsequently, only genus and species features presenting significantly different profiles in both DEseq2 and pooling RF models in 8 strategies are shown in a heat map. Different enrichment patterns of the microbiota were hierarchically clustered using the Pearson correlation, as presented in rows. Alpha diversity regarding this part was performed based on relative abundance-transformed feature counts.

Statistical analysis. The Mann-Whitney test was applied to compute the paired-sample difference and significance, and the Kruskal-Wallis rank sum test was used for multiple samples. DEseq2 (46) was chosen to conduct differential feature abundance analyses with cohort information added to adjust the batch effect. Log_2-transformed fold changes and adjusted *P* values served as factors for downstream screening. The ade4 (47) R package was used to compute 95% of the inertia in the PCoA modules for each group. The 95% confidence interval of the ROC was calculated with 2,000 stratified bootstrap replicates, and DeLong's test was conducted for two ROC curves using the pROC R package (48). The Mantel test was applied to compute the similarity and significance between matrices using the two-sided method implemented in the ade4 package (47).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.9 MB.

FIG S2, PDF file, 2 MB.

FIG S3, PDF file, 0.2 MB.

FIG S4, PDF file, 1 MB.

FIG S5, PDF file, 1.7 MB.

FIG S6, PDF file, 0.9 MB.

TABLE S1, XLS file, 0.1 MB.

TABLE S2, XLS file, 0.2 MB.

TABLE S3, XLS file, 0.1 MB.

TABLE S4, XLS file, 0.1 MB.

ACKNOWLEDGMENTS

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

We thank Burkhardt Flemer for sharing his data with us on our request. We thank other scientists for sharing their data sets in their publications.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394–424. https://doi.org/10.3322/caac.21492.
2. Siegel RL, Miller KD, Jemal A. 2019. Cancer statistics, 2019. CA Cancer J Clin 69:7–34. https://doi.org/10.3322/caac.21551.
3. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, Sunagawa S, Coelho LP, Schrotz-King P, Vogtmann E, Habermann N, Niméus E, Thomas AM, Manghi P, Gandini S, Serrano D, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Waldron L, Naccarati A, Segata N, Sinha R, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 25:679–689. https://doi.org/10.1038/s41591-019-0406-6.
4. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, Gandini S, Serrano D, Tarallo S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A, Segata N. 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 25:667–678. https://doi.org/10.1038/s41591-019-0405-7.
5. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Tran Van Nhieu J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, von Knebel Doeberitz M, Sobhani I, Bork P. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 10:766. https://doi.org/10.15252/msb.20145645.
6. Shah MS, DeSantis T, Yamal JM, Weir T, Ryan EP, Cope JL, Hollister EB. 2018. Re-purposing 16S rRNA gene sequence data from within case paired tumor biopsy and tumor-adjacent biopsy or fecal samples to identify microbial markers for colorectal cancer. PLoS One 13:e0207002. https://doi.org/10.1371/journal.pone.0207002.
7. Liu CJ, Zhang YL, Shang Y, Wu B, Yang EN, Luo YY, Li XR. 2019. Intestinal bacteria detected in cancer and adjacent tissue from patients with colorectal cancer. Oncol Lett 17:1115–1127. https://doi.org/10.3892/ol.2018.9714.
8. Flemer B, Lynch DB, Brown JMR, Jeffery IB, Ryan FJ, Claesson MJ, O'Riordain M, Shanahan F, O'Toole PW. 2017. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut 66:633–643. https://doi.org/10.1136/gutjnl-2015-309595.
9. Zackular JP, Rogers MAM, Ruffin MT, Schloss PD. 2014. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res 7:1112–1121. https://doi.org/10.1158/1940-6207.CAPR-14-0129.
10. Baxter NT, Ruffin MT, IV, Rogers MAM, Schloss PD. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med 8:37. https://doi.org/10.1186/s13073-016-0290-3.
11. Shah MS, DeSantis TZ, Weinmaier T, McMurdie PJ, Cope JL, Altrichter A, Yamal JM, Hollister EB. 2018. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. Gut 67:882–891. https://doi.org/10.1136/gutjnl-2016-313189.
12. Sze MA, Schloss PD. 2018. Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. mBio 9:e02076-18. https://doi.org/10.1128/mBio.02076-18.
13. He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, Chen MX, Chen ZH, Ji GY, Zheng ZDX, Mujagond P, Chen XJ, Rong ZH, Chen P, Lyu LY, Wang X, Wu CB, Yu N, Xu YJ, Yin J, Raes J, Knight R, Ma WJ, Zhou HW. 2018. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 24:1532–1535. https://doi.org/10.1038/s41591-018-0164-x.
14. Zoetendal EG, Von Wright A, Vilpponen-Salmela T, Ben-Amor K, Akkermans ADL, De Vos WM. 2002. Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. Appl Environ Microbiol 68:3401–3407. https://doi.org/10.1128/AEM.68.7.3401-3407.2002.
15. Chen W, Liu F, Ling Z, Tong X, Xiang C. 2012. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PLoS One 7:e39743. https://doi.org/10.1371/journal.pone.0039743.

16. Li S, Peppelenbosch MP, Smits R. 2019. Bacterial biofilms as a potential contributor to mucinous colorectal cancer formation. Biochim Biophys Acta 1872:74–79. https://doi.org/10.1016/j.bbcan.2019.05.009.
17. Yazici C, Wolf PG, Kim H, Cross TWL, Vermillion K, Carroll T, Augustus GJ, Mutlu E, Tussing-Humphreys L, Braunschweig C, Xicola RM, Jung B, Llor X, Ellis NA, Gaskins HR. 2017. Race-dependent association of sulfidogenic bacteria with colorectal cancer. Gut 66:1983–1994. https://doi.org/10.1136/gutjnl-2016-313321.
18. Mira-Pascual L, Cabrera-Rubio R, Ocon S, Costales P, Parra A, Suarez A, Moris F, Rodrigo L, Mira A, Collado MC. 2015. Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers. J Gastroenterol 50:167–179. https://doi.org/10.1007/s00535-014-0963-x.
19. Thomas AM, Jesus EC, Lopes A, Aguiar S, Begnami MD, Rocha RM, Carpinetti PA, Camargo AA, Hoffmann C, Freitas HC, Silva IT, Nunes DN, Setubal JC, Dias-Neto E. 2016. Tissue-associated bacterial alterations in rectal carcinoma patients revealed by 16S rRNA community profiling. Front Cell Infect Microbiol 6:179. https://doi.org/10.3389/fcimb.2016.00179.
20. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20:257. https://doi.org/10.1186/s13059-019-1891-0.
21. Leslie A, Carey FA, Pratt NR, Steele RJC. 2002. The colorectal adenoma-carcinoma sequence. Br J Surg 89:845–860. https://doi.org/10.1046/j.1365-2168.2002.02120.x.
22. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. PLoS Comput Biol 8:e1002687. https://doi.org/10.1371/journal.pcbi.1002687.
23. Nakatsu G, Li X, Zhou H, Sheng J, Wong SH, Wu WKK, Ng SC, Tsoi H, Dong Y, Zhang N, He Y, Kang Q, Cao L, Wang K, Zhang J, Liang Q, Yu J, Sung JJY. 2015. Gut mucosal microbiome across stages of colorectal carcinogenesis. Nat Commun 6:8727. https://doi.org/10.1038/ncomms9727.
24. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. 2013. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. Cell Host Microbe 14:195–206. https://doi.org/10.1016/j.chom.2013.07.012.
25. Dai Z, Coker OO, Nakatsu G, Wu WKK, Zhao L, Chen Z, Chan FKL, Kristiansen K, Sung JJY, Wong SH, Yu J. 2018. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome 6:70. https://doi.org/10.1186/s40168-018-0451-2.
26. Watt E, Gemmell MR, Berry S, Glaire M, Farquharson F, Louis P, Murray GI, El-Omar E, Hold GL. 2016. Extending colonic mucosal microbiome analysis—assessment of colonic lavage as a proxy for endoscopic colonic biopsies. Microbiome 4:61. https://doi.org/10.1186/s40168-016-0207-9.
27. Suez J, Zmora N, Zilberman-Schapira G, Mor U, Dori-Bachash M, Bashiardes S, Zur M, Regev-Lehavi D, Ben-Zeev Brik R, Federici S, Horn M, Cohen Y, Moor AE, Zeevi D, Korem T, Kotler E, Harmelin A, Itzkovitz S, Maharshak N, Shibolet O, Pevsner-Fischer M, Shapiro H, Sharon I, Halpern Z, Segal E, Elinav E. 2018. Post-antibiotic gut mucosal microbiome reconstitution is impaired by probiotics and improved by autologous FMT. Cell 174:1406–1423.e16. https://doi.org/10.1016/j.cell.2018.08.047.
28. Stintzi A, Abujamel T, Butcher J, Li J, Mack D, Manoogian J, Mottawea W. 2019. The mucosal-luminal interface: an ideal sample to study the mucosa-associated microbiota and the intestinal microbial biogeography. Pediatr Res 85:895–903. https://doi.org/10.1038/s41390-019-0326-7.
29. Budding AE, Grasman ME, Eck A, Bogaards JA, Vandenbroucke-Grauls CMJE, Van Bodegraven AA, Savelkoul PHM. 2014. Rectal swabs for analysis of the intestinal microbiota. PLoS One 9:e101344. https://doi.org/10.1371/journal.pone.0101344.
30. Bassis CM, Moore NM, Lolans K, Seekatz AM, Weinstein RA, Young VB, Hayden MK, CDC Prevention Epicenters Program. 2017. Comparison of stool versus rectal swab samples and storage conditions on bacterial community profiles. BMC Microbiol 17:78. https://doi.org/10.1186/s12866-017-0983-9.
31. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
32. Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, Hurley E, O'Riordain M, Shanahan F, O'Toole PW. 2018. The oral microbiota in colorectal cancer is distinctive and predictive. Gut 67:1454–1463. https://doi.org/10.1136/gutjnl-2017-314814.
33. Burns MB, Lynch J, Starr TK, Knights D, Blekhman R. 2015. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. Genome Med 7:55. https://doi.org/10.1186/s13073-015-0177-8.
34. Drewes JL, White JR, Dejea CM, Fathi P, Iyadorai T, Vadivelu J, Roslani AC, Wick EC, Mongodin EF, Loke MF, Thulasi K, Gan HM, Goh KL, Chong HY, Kumar S, Wanyiri JW, Sears CL. 2017. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. NPJ Biofilms Microbiomes 3:34. https://doi.org/10.1038/s41522-017-0040-3.
35. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, Baselga J, Liu C, Shivdasani RA, Ogino S, Birren BW, Huttenhower C, Garrett WS, Meyerson M. 2012. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res 22:292–298. https://doi.org/10.1101/gr.126573.111.
36. Dejea CM, Wick EC, Hechenbleikner EM, White JR, Mark Welch JL, Rossetti BJ, Peterson SN, Snesrud EC, Borisy GG, Lazarev M, Stein E, Vadivelu J, Roslani AC, Malik AA, Wanyiri JW, Goh KL, Thevambiga I, Fu K, Wan F, Llosa N, Housseau F, Romans K, Wu X, McAllister FM, Wu S, Vogelstein B, Kinzler KW, Pardoll DM, Sears CL. 2014. Microbiota organization is a distinct feature of proximal colorectal cancers. Proc Natl Acad Sci U S A 111:18321–18326. https://doi.org/10.1073/pnas.1406199111.
37. Geng J, Fan H, Tang X, Zhai H, Zhang Z. 2013. Diversified pattern of the human colorectal cancer microbiome. Gut Pathog 5:2. https://doi.org/10.1186/1757-4749-5-2.
38. Lu Y, Chen J, Zheng J, Hu G, Wang J, Huang C, Lou L, Wang X, Zeng Y. 2016. Mucosal adherent bacterial dysbiosis in patients with colorectal adenomas. Sci Rep 6:26337. https://doi.org/10.1038/srep26337.
39. Cong J, Zhu H, Liu D, Li T, Zhang C, Zhu J, Lv H, Liu K, Hao C, Tian Z, Zhang J, Zhang X. 2018. A pilot study: changes of gut microbiota in post-surgery colorectal cancer patients. Front Microbiol 9:2777. https://doi.org/10.3389/fmicb.2018.02777.
40. Gao R, Kong C, Huang L, Li H, Qu X, Liu Z, Lan P, Wang J, Qin H. 2017. Mucosa-associated microbiota signature in colorectal cancer. Eur J Clin Microbiol Infect Dis 36:2073–2083. https://doi.org/10.1007/s10096-017-3026-4.
41. Nakatsu G, Zhou H, Wu WKK, Wong SH, Coker OO, Dai Z, Li X, Szeto CH, Sugimura N, Lam TYT, Yu ACS, Wang X, Chen Z, Wong MCS, Ng SC, Chan MTV, Chan PKS, Chan FKL, Sung JJY, Yu J. 2018. Alterations in enteric virome are associated with colorectal cancer and survival outcomes. Gastroenterology 155:529–541.e5. https://doi.org/10.1053/j.gastro.2018.04.018.
42. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9:811–814. https://doi.org/10.1038/nmeth.2066.
43. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336. https://doi.org/10.1038/nmeth.f.303.
44. Holmes I, Harris K, Quince C. 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One 7:e30126. https://doi.org/10.1371/journal.pone.0030126.
45. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541. https://doi.org/10.1128/AEM.01541-09.
46. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8.
47. Dray S, Dufour A-B. 2007. The ade4 package: implementing the duality diagram for ecologists. J Stat Softw 22:1–20.
48. Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, Fouda C, Bassem H, Mueller M, Lisacek F, Puybasset L, Sanchez J-C. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77. https://doi.org/10.1186/1471-2105-12-77.