Identification of microbial markers across populations in early detection of colorectal cancer

Yuanqi Wu[1, $], Na Jiao[2, $], Ruixin Zhu[1, *], Yida Zhang[3], Dingfeng Wu[1], An-Jun Wang[4], Sa Fang[1], Liwen Tao[1], Yichen Li[2], Sijing Cheng[2, 5], Xiaosheng He[2], Ping Lan[2, 5], Chuan Tian[1, *], Ning-Ning Liu[4, *], Lixin Zhu[2, 6, *]

1. Department of Gastroenterology, The Shanghai Tenth People's Hospital, Department of Bioinformatics, School of Life Sciences and Technology, Tongji University, Shanghai 200092, P.R.China.
2. Guangdong Institute of Gastroenterology, Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, the Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, P.R.China
3. Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, United States
4. State Key Laboratory of Oncogenes and Related Genes, Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, P.R.China
5. School of Medicine, Sun Yat-sen University, Guangzhou 510006/Shenzhen 518107, P.R.China
6. Genome, Environment and Microbiome Community of Excellence, the State University of New York at Buffalo, Buffalo, New York 14214, United States
$ Equal contribution, * Corresponding authors

Corresponding authors:

Lixin Zhu (zhulx6@mail.sysu.edu.cn)
Guangdong Institute of Gastroenterology, Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Department of Colorectal Surgery, the Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, P.R. China.
Tel: 86-199-4625-6235

Ning-Ning Liu (liuningning@shsmu.edu.cn)
State Key Laboratory of Oncogenes and Related Genes, Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, P.R.China
Tel: 86-21-5359-4658

Chuan Tian (tianchuan1015@gmail.com)
Department of Gastroenterology, The Shanghai Tenth People's Hospital, Department of Bioinformatics, School of Life Sciences and Technology, Tongji University, Shanghai 200092, P.R.China.
Currently at Eli Lilly and Company, 10290 Campus Point Dr, San Diego, CA 92121.
Tel: 1-631-682-4501

Ruixin Zhu (rxzhu@tongji.edu.cn)
Department of Gastroenterology, The Shanghai Tenth People's Hospital, Department of Bioinformatics, School of Life Sciences and Technology, Tongji University, Shanghai 200092, P.R.China.
Tel: 86-21-6598-1041

Abstract

Several studies have investigated the association between microbial and colorectal cancer (CRC). However, the replicable markers for early stage adenoma diagnosis across multiple populations remain elusive. Here, a meta-analysis of six studies, comprising a total of 1057 fecal samples, was performed to identify candidate markers. By adjusting the potential confounders, 11 and 26 markers (*P*<0.05) were identified and separately applied into constructing Random Forest classifier models to discriminate adenoma from control, and adenoma from CRC, achieving robust diagnostic accuracy with AUC = 0.80 and 0.89, respectively. Moreover, these markers demonstrated high diagnostic accuracy in independent validation cohorts. Pooled functional analysis and targeted qRT-PCR based genetic profiles reveal that the altered microbiome triggers different pathways of ADP-heptose and menaquinone biosynthesis (*P*<0.05) in adenoma vs. control and adenoma vs. CRC sequences respectively. The combined analysis of heterogeneous studies confirm adenoma-specific but universal markers across multi-populations, which improves early diagnosis and prompt treatment of CRC.

Keywords: colorectal cancer, adenoma, gut microbiota, meta-analysis, early diagnosis

Introduction

Colorectal cancer (CRC) is one of the most common cancer with an overall high mortality rate. According to the report of the International Agency for Research on Cancer (IARC), there were over 1,800,000 new CRC cases and over 860,000 deaths in 2018(1). And CRC accounted for approximately 10% of all new cancer cases globally(2). It is estimated that the national expenditures in the United States on cancer care, specifically colorectal cancer, were about 16.63 billion dollars in 2018(3), and the CRC burden is continuously growing over years. Colorectal adenomas are recognized as precursors for the majority of CRC(2). The early detection of CRC at precancerous-stage adenoma has increased the 5-year relative survival rate to about 90%, significantly facilitating early decision making, alleviating the incidence of CRC and reducing economic burden(2, 4).

Gut microbiome is a novel stool-based non-invasive biomarker for metabolic diseases and cancers(5, 6). Many studies have reported that the gut microbiome is an important aetiological element in the initiation and progression of CRC(4, 7) and identified some fecal microbial markers of CRC(8-10). However, there is limited knowledge on whether these biomarkers could more precisely detect early-stage of CRC, adenomas. And this cognitive gap needs to be filled with more intellectual efforts. Furthermore, current knowledge of the associations between microbiome and biomarkers for colorectal adenoma early-detection is poor as well. Only a few studies have investigated the microbial alterations in colorectal adenoma(4, 7, 11-13). However, a substantial variation exists among microbial makers in these studies, and its cause could be various biological factors influencing gut microbiome composition and inconsistent processing of microbial sequencing data.

Meta-analysis offers a set of tools that is powerful, informative and unbiased to improve the robustness of microbiome alterations and reduce the noise of biological and technical confounders so that consistent alterations across multiple studies could be identified. Recently, several meta-analysis of multi-studies have identified universal microbial markers across multiple diseases, such as CRC(11, 13-15), obese(16), Inflammatory bowel disease (IBD)(17), via 16S rRNA sequencing or whole metagenome shotgun sequencing (WMS) technique. However, previous researches based on meta-analysis(11, 13) still could not identify universal stool-based microbial markers for colorectal cancer across multiple cohorts (Supplementary Note 1). Additionally, the commonly used non-invasive stool-based screening test, Faecal Immunochemical Test (FIT), has drawbacks such as poor sensitivity to early and advanced adenoma (7.6% and 38%, respectively)(18). Therefore, it is urgent to explore and identify novel stool-based microbial markers that could more precisely and efficiently diagnose colorectal adenoma and its various stages.

Here, we presented a meta-analysis study, aiming to identify a series of markers that enable distinguishing adenoma from healthy control or CRC with high accuracy across multiple cohorts. We included fecal 16S rRNA sequencing studies considering that 16S rRNA gene-based profiles are more closely matching the “real community”(19). We then investigated the potential mechanisms of the disordered microbiome in colorectal adenomas, which may provide biological insights and therapeutical strategies to detect early syndromes and alleviate symptoms of CRC.

Results

Characteristics of the datasets in meta-analysis

In this study, we investigated 16S rRNA sequencing data from four studies to measure the gut microbiome changes as CRC progresses (from control to adenoma to cancer) and to identify the biomarkers specific to adenoma. In total, we collected 307 samples from colorectal adenoma patients, 217 from CRC subjects and 252 samples as control. The demographic information was listed in detail in Table 1. All samples were sequenced at sufficient depth, with average counts of 85637 in each sample. Consistent processing was performed for all raw sequencing data on the QIIME2 platform.

Table 1 Characteristics of the large-scale adenoma datasets included in this study

Study	Group(N)*	Age (average±s.d.)#	BMI (average±s.d.)	Sex F(%)/M(%)†	Country
CA(12)	Control(30)	55.27±9.22	26.73±5.19	63.30/36.70	American Canadian
	Adenoma(30)	61.30 ±11.15	27.40 ±4.45	60.00/40.00	
	Cancer(30)	59.40 ±10.99	30.59±7.18	70.00/30.00	
FR(20)	Control(50)	62.32±8.98	24.66±4.69	52.00/48.00	France
	Adenoma(38)	62.29±8.51	27.40 ±4.45	28.90/71.10	
	Cancer(41)	65.51±10.51	30.59±7.18	41.50/58.50	
US1(21)	Adenoma(41)	62.34±9.01	26.37±4.28	34.10/65.90	American
	Cancer(26)	61.65±12.89	28.63±7.19	42.30/57.70	
US2(22)	Control(172)	54.29±9.93	26.69±5.33	64.50/35.50	American
	Adenoma(198)	63.35±11.47	26.27±4.73	40.40/59.60	
	Cancer(41)	63.78±12.89	28.89±7.25	43.30/56.70	
Total:	Control(252)	56.00±10.14	26.48±5.25	61.90/38.10	
	Adenoma(307)	62.89±10.80	26.21±4.80	38.89/61.11	
	Cancer(217)	63.25±12.28	28.30±7.23	41.01/58.99	

*** Number of samples**

# Standard deviation;

† The ratio of percentage of female and male

Identification of the potential confounder in meta-analysis

Since differences existed among these studies in both technical and biological aspects, we first investigated the potential confounders. The variances explained by disease status for each ASV were calculated to quantify the effects of potential confounders (see method confounder analysis) (Supplementary Fig. 1, 2). This analysis revealed that the factor 'study' had a predominant impact on microbial composition (Fig. 1a and Supplementary Fig. 1). Additionally, the microbial alpha and beta diversity also supported that the heterogeneity of studies had a more significant impact on microbial composition than disease status (Fig. 1b and Supplementary Fig. 3). Therefore, we treated 'study' as a blocking factor in the subsequent analysis.

Alterations of gut microbial composition in colorectal adenoma

At the phylum level, the gut microbiota was dominated by members of Firmicutes and Bacteroidetes, followed by Proteobacteria, Actinobacteria, Verrucomicrobia, Tenericutes and Fusobacteria in healthy controls, adenomas and CRC. These dominant phyla were similar to those reported in previous studies on gut microbiota(20). Furthermore, the phylum Fusobacteria, the most CRC-associated bacteria as reported(23), were observed with significantly decreased abundance in adenoma compared to that in cancer, while there was no significant difference between adenoma patients and controls (Fig. 1c).

At the ASV level, significant alterations across studies were observed among different disease status. In the comparison of gut communities between controls and patients with adenoma, 43 ASVs were identified with distinguishable abundances (Supplementary Note 2). Moreover, we also identified 114 differentially abundant ASVs between adenoma and cancer (Supplementary Note 3).

Additionally, pathogenic bacteria with increased abundance were detected in adenoma or cancer compared with control. For instance, *Parvimonas* genus was enriched in adenoma compared with controls while *Fusobacterium*, *Porphyromonas*, *Peptostreptococcus*, *Parvimonas*, and *Escherichia-Shigella* genus were enriched in cancer compared with adenoma. Particularly, *Fusobacterium*, *Porphyromonas*, *Parvimonas* and *Peptostreptococcus* were identified as oral pathogens associated with CRC(17, 24). Notably, there were only 9 common differential ASVs between healthy controls versus adenoma and adenoma versus cancer, which could be further classified into Ruminococcaceae, Lachnospiraceae, Family XI and Veillonellaceae family (Fig. 1d). The two sets of differential ASVs with a Jaccard distance of 0.939 indicate that the microbiota has a remarkable difference between adenoma and control or cancer.

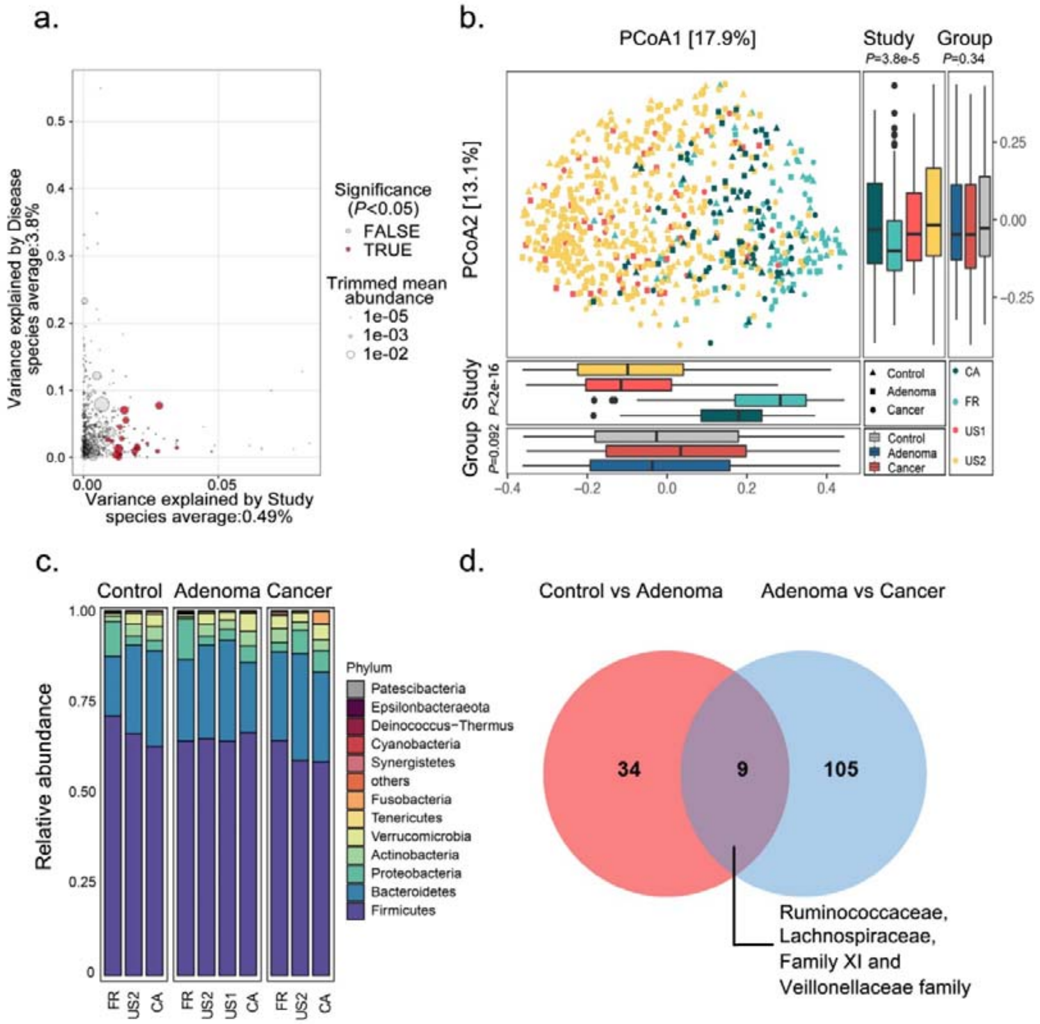

Fig. 1 Alterations of gut microbial composition in different disease status accounting for study heterogeneity. a, Variance explained by disease status (adenoma versus cancer) is plotted against variance explained by study effects for individual ASVs. The significantly differential ASVs are colored in red and the dot size is proportional to the abundance of each ASV. b, Principal coordinate analysis of samples from all four studies based on Bray-Curtis distance; the study is color-coded and the group (control, adenoma and cancer) is indicated by different shapes. The upper-right and the bottom-left boxplots illustrate that samples projected onto the first two principal coordinates broken down by study and disease status, respectively. *P* values were calculated with a Kruskal–Wallis test for study and group. All boxplots represent 25th–75th percentile of the distribution; the median is shown in thick line at the middle of the box; the whiskers extend up to values within 1.5 times, and outliers are represented as dots. c, Relative proportions of bacterial phyla in healthy controls, adenomas and CRC across four different studies. d, Venn diagram shows the overlap of differential ASVs between adenomas and healthy controls or CRC.

Microbial classification models for colorectal adenoma

Next, we constructed RF models by pooling all samples to select features capable of distinguishing adenoma from control and cancer. Besides using differential ASVs as key metrics, alpha diversity indices including Shannon Index, Simpson Index and Observed ASVs, and three patient metadata, age, gender and BMI were also included in model building. To obtain the best performing models and most important features, an IEF step was further applied.

A robust RF model was constructed with a core set of important features, including 8 differential ASVs (as biomarkers) together with age, gender and BMI, which were proved to have the best capability to distinguish control subjects from patients with adenoma (AUC = 0.80) (Fig. 2a, c and Supplementary Table 3). Among these, the ASV assigned as *Christensenellaceae R-7 group sp*. was the highest-ranking biomarker (Fig. 2a). The biomarkers also included the increased abundance ASVs of *[Eubacterium] coprostanoligenes group, Ruminiclostridium 9 sp.*, *Christensenellaceae R-7 group sp.*, *Ruminococcaceae UCG-005 sp.* and *Veillonella parvula* as well as the decreased abundance of *Rothia dentocariosa* and *Aminipila butyrica* in adenoma (Supplementary Fig. 4).

Similarly, the best performance of the RF model for distinguishing adenoma from cancer is 0.89 (AUC). The RF model was built with 24 ASVs together with age and BMI (Fig. 2b, d and Supplementary Table 4). Among these biomarkers, the ASV belonging to *Streptococcus thermophilus TH1435* was the top-ranking biomarker(Fig. 2b). The following ASVs were assigned as *Parvimonas micra*, *Bacteroides dorei*, *[Clostridium] scindens*, *Erysipelatoclostridium ramosum*, *Blautia sp.*, *[Eubacterium] coprostanoligenes group sp.* and *Lachnospira pectinoschiza* (Fig. 2b). The *[Clostridium] scindens* was significantly (P < 0.001) enriched in cancer compared with adenoma with a generalized fold change of 0.49. Additionally, the abundance of *Blautia sp.*, *Hungatella hathewayi WAL-18680* and *Eubacterium ruminantium* were gradually increased while *Streptococcus thermophilus TH1435*, *Erysipelatoclostridium ramosum*, *[Eubacterium] ventriosum group sp.* and *Roseburia intestinalis* were gradually decreased during CRC carcinogenesis (Supplementary Fig. 5). In the two models, age was ranked as the top and third predictor in the testing phase, respectively. In the two sets of biomarkers, there was only one common ASV classified as *Eubacterium ruminantium*.

Moreover, we also identified that a core set of 34 ASVs, together with age, gender and BMI, collectively had the highest capability to distinguish control from cancer (AUC=0.93) (Supplementary Fig. 6, Supplementary Note 4). It is worth noting that there was no common ASV in the two sets of biomarkers between healthy controls and adenomas or CRC(Supplementary Fig. 7). Thus these results highlighted that microbial markers aimed to detect CRC are specific and exclusive, and would not be used as optimal diagnosis of adenoma.

Co-occurrence and clustering analysis of microbiota in different states

Through the co-occurrence network of differential ASVs, our results suggested that most of the identified biomarkers have functional importance in the network (Supplementary Note 5). To gain further insight, we analyzed metagenomes of patients in adenoma and control. Co-occurrences analysis demonstrated four clusters of biomarkers with distinct taxonomic composition (Supplementary Fig. 9a). These clusters are not tightly associated with patient characteristics such as Age, Sex and BMI (Supplementary Fig. 10a), revealing that the adenoma-associated microbiota closely resembles that of the healthy control. These results further proved the high detection accuracy (AUC of 0.8) and overall success of merely using 11 important features to distinguish control from adenoma.

Moreover, we also explored the CRC patient metagenomes for co-occurrences among a panel of 24 biomarkers and yielded three clusters (Supplementary Fig. 9b). Cluster 2 demonstrated strong taxonomic consistency, which was primarily comprised of members following Clostridiales order. In contrast, the other two clusters exhibited heterogeneous taxonomy, with cluster 1 containing high-ranking biomarkers and cluster 3 assorting together the species that highly prevailed in CRC individuals. We then investigated the association between these three clusters and various tumor characteristics. Clostridiales cluster 2 is significantly enriched in male CRC patients. Besides, both cluster 1 and cluster 3 show a slight tendency toward late-stage CRC (containing stages 3 and 4 according to the American Joint Committee on Cancer), and this tendency is significant for cluster 3. Associations with patient age and BMI are weaker and not significant (Supplementary Fig. 10b). Based on these results, it can be deduced that the adenoma-associated microbiota differs from that of CRC. To consider the impact of using different studies, all of these tests were adjusted by blocking for "study" (see method co-occurrence and clustering analysis).

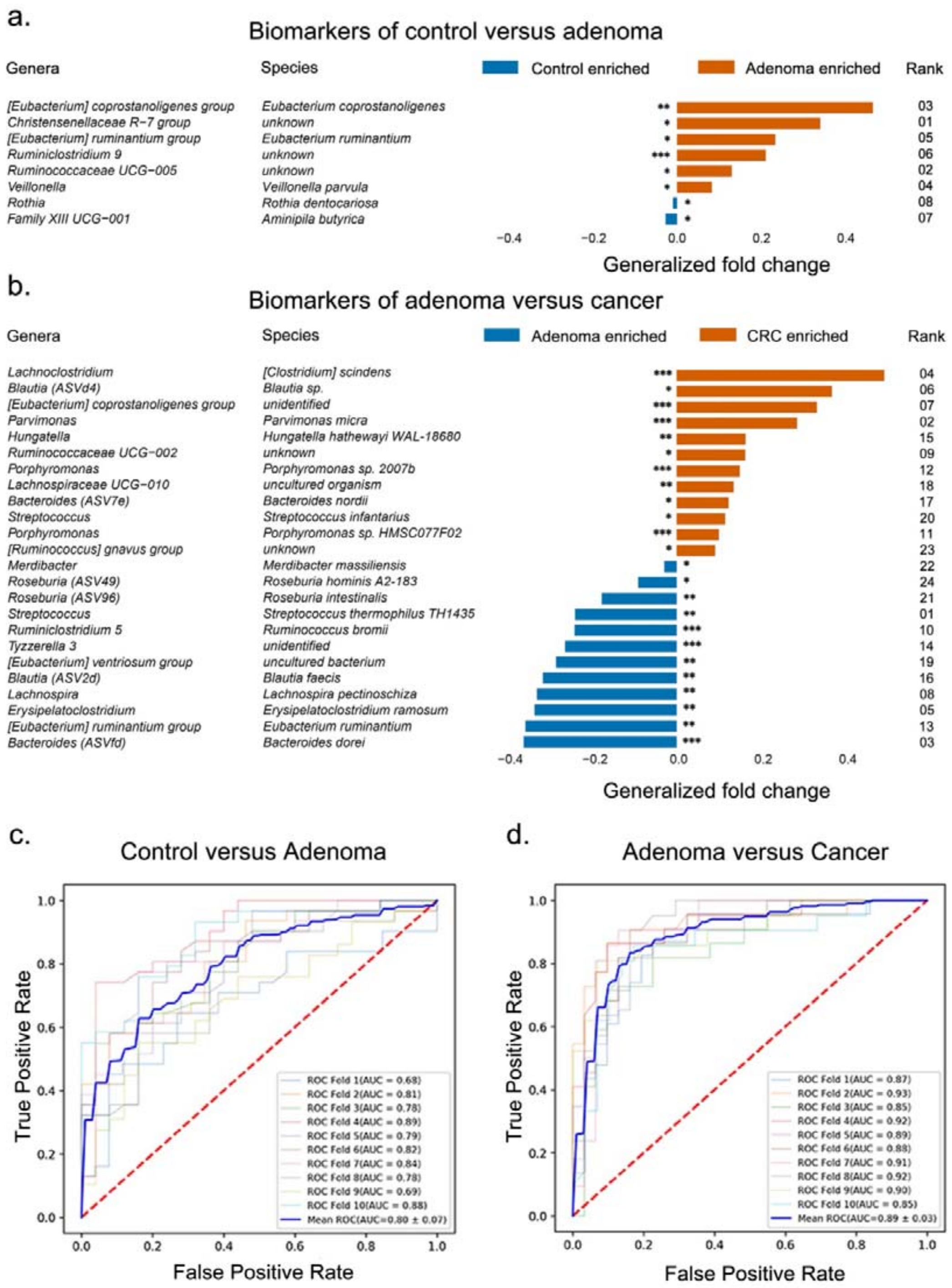

Fig. 2 Performance of discriminating adenoma from control or cancer using important features. a, b, The important biomarkers identified to construct RF model for discriminating adenoma from control (a) and CRC (b). The rank in (a) and (b) means the order of feature importance in the RF model; *: $P <$ 0.05, **: $P < 0.01$and **: $P < 0.001$. c, d, The AUC of the optimized models constructed with biomarkers and metadata of Control versus Adenoma (c) and Adenoma versus Cancer (d). Mean ROC in (c) and (d): the average AUC from tenfold cross-validation.

Validation of the colorectal adenoma classifiers

To test whether the selected important features are universal and robust across multiple studies, we performed study-to-study transfer validation and LODO validation on the entire samples. In study-to-study transfer validation, the average AUC for healthy controls versus adenoma model was 0.64 and the AUCs of all datasets ranged from 0.52 and 0.86, which maintained the diagnostic accuracy of within-study (Fig. 3a). Notably, the US2 study serves as a better training set than other studies through exhibiting relatively higher testing AUCs. The reason could be that the US2 study has a larger data set that is beneficial to develop accurate classifiers. Moreover, we also compare the diagnostic performance of selected important features with FIT, which illustrates improved adenoma diagnostic ability by combining with non-invasive clinical screening tests (Supplementary Note 6). Additionally, in the LODO analysis, the AUC values of control versus adenoma models range from 0.61 to 0.87 with an average 0.72, which is superior to study-to-study transfer validation owing to using large datasets (Fig. 3a). This reveals that more training samples would in principle improve the robustness of classifiers.

Similar results were observed in the adenoma versus cancer model (Fig. 3b). The average AUC of study-to-study transfer validation is 0.76 and the AUCs of all datasets range from 0.59 to 0.93. Besides, the classification accuracy of the CA within-study is relatively low. The possible reason is that the CA study was small-sized and subjects came from different countries, which indicates that classifiers may be geographic region-specific when dataset is limited. This result reinforced Zhou's research that region variation limits the usefulness of disease modelling(25). Moreover, the AUC values are also elevated in the LODO analysis, ranging from 0.86 to 0.95 with an average 0.89 (Fig. 3b). We notice that the classifiers performed better in adenoma versus cancer than that in control versus adenoma, which reinforced previous findings that the adenoma-associated stool microbiome closely resembled that of the health status(7, 11, 20).

To determine the maximum subset of important features required to provide comparable accuracy on validation studies and methods, we analyzed sets of features including all ASVs (all), differentially abundant ASVs (control versus adenoma:43, adenoma versus cancer:114), all important features (control versus adenoma:11, adenoma versus cancer:26) and reduced important features according to the feature ranks of the RF classifiers. In both study-to-study transfer validation (Fig. 3c, d) and LODO validation (Supplementary Fig. 12a, b), as the number of important features increases, the average AUC increases and reaches maximum when all the important features are included for all studies except the CA study. This may also be owing to the characteristics of small-sized and geographic heterogeneity in the CA study. As we continued to add more ASVs, especially the ones not part of important features among disease status, the average AUC of cross-validation decreases. Therefore, this result further confirmed that the sets of selected important features contributed to the accuracy of classifiers.

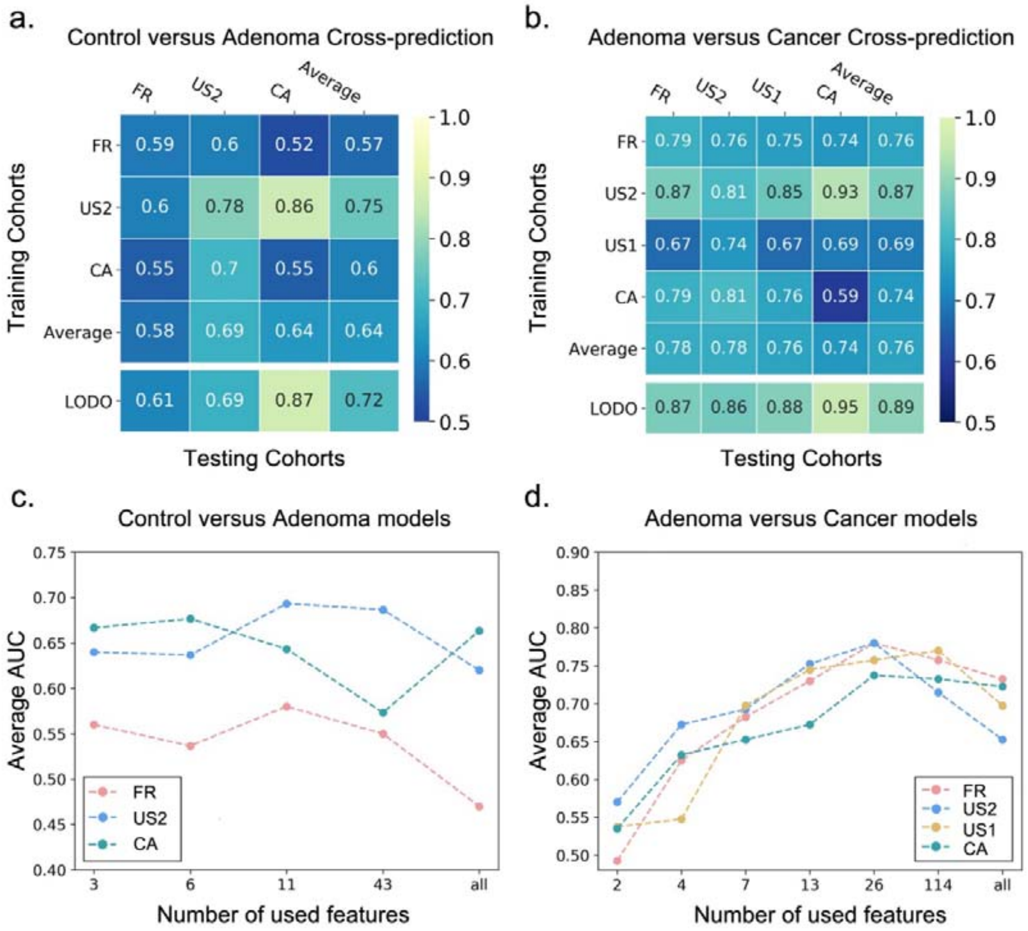

Fig. 3 Prediction performance of important features across studies and identification of minimal features for detecting adenoma. a, b, Cross-prediction matrix depicting prediction values for differentiating adenoma from control (a) and CRC (b) as AUC obtained using important features. Values on the diagonal refer to the results of within-cohort validation; Off-diagonal values refer to the AUC values obtained from cross-cohort validation, which training the classifier on the study of the corresponding row and applying it to the study of the corresponding column; The LODO values refer to the performances obtained by training the classifier using all but the study of the corresponding column and applying it to the study of the corresponding column (see methods). c, d, Average AUC of study-to-study transfer validation classifiers for control versus adenoma (c) and adenoma versus cancer (d) at different sets of features. The x-axis in (c) and (d) indicate different sets of features: All (c-d): all ASVs; 43 (c) and 114 (d): differentially abundant ASVs; 11 (c) and 26 (d): all important features and other top-ranking important features. The different studies were indicated in different colors.

Validation of colorectal adenoma markers in independent cohorts

To further validate our meta-analysis results, two additional independent cohorts from America (Validation Cohort1) and China (Validation Cohort2) were incorporated into this study. The validation cohort1 is comprised of 70 controls and 102 adenoma patients, while there were 57 adenoma patients and 52 CRC patients in the validation cohort2(Supplementary Table 8). The independent predictive RF model was confirmed to be relatively accurate on the two new cohorts, with an AUC of 0.73 and 0.83 for distinguishing adenoma from controls or cancer, respectively (Supplementary Fig. 13a, b). Although the validation cohort2 was lack of patient metadata, it obtained a relatively high AUC, indicating that the gut microbial biomarkers could distinguish adenomas from CRC precisely. Additionally, *Ruminococcaceae UCG-005 sp.* and *Christensenellaceae R-7 group sp.* were confirmed as the top-ranking biomarkers between controls and adenoma patients in validation cohort1. Furthermore, *Parvimonas micra*, *Streptococcus thermophilus TH1435* and *Bacteroides dorei* were confirmed as the three top-ranking biomarkers for distinguishing between adenoma and CRC patients in validation cohort2.

The specificity of colorectal adenoma predictive models

After evaluating the accuracy of the above colorectal adenoma predictive models on different cohorts, we further validated the specificity of colorectal adenoma related important features in other potentially microbiome-linked diseases. Five microbiome-linked diseases including NAFLD, T2D, CD, UC and IBS were considered in this analysis(Supplementary Table 8). We randomly drew samples from each disease and the control of these non-CRC studies and added them to the control class of the validation cohort1. By comparing AUC scores between adding non-CRC cases and adding the corresponded external controls, we found a small decrease (ranging from 1% to 4%) in prediction accuracy for the non-CRC group(Supplementary Fig. 14). When adding both control and case samples of the IBD study, the AUCs decreased more than other studies, which may be caused by lack of several key features including BMI and two biomarkers. Taken together, these results indicated that the colorectal adenoma-specific model might not be necessarily applicable to other microbiome-associated diseases.

Microbial functional changes in colorectal adenoma

We investigated the microbial-based functional alterations for multiple different disease status. There are 27 differential pathways between control and adenoma (Supplementary Table 9) and 41 differential pathways between adenoma and cancer (Supplementary Table 10) consistently detected across studies. A total of 64 differential pathways (4 pathways were overlapped) were clustered based on their generalized fold change scores. (Fig. 4, Supplementary Note 7).

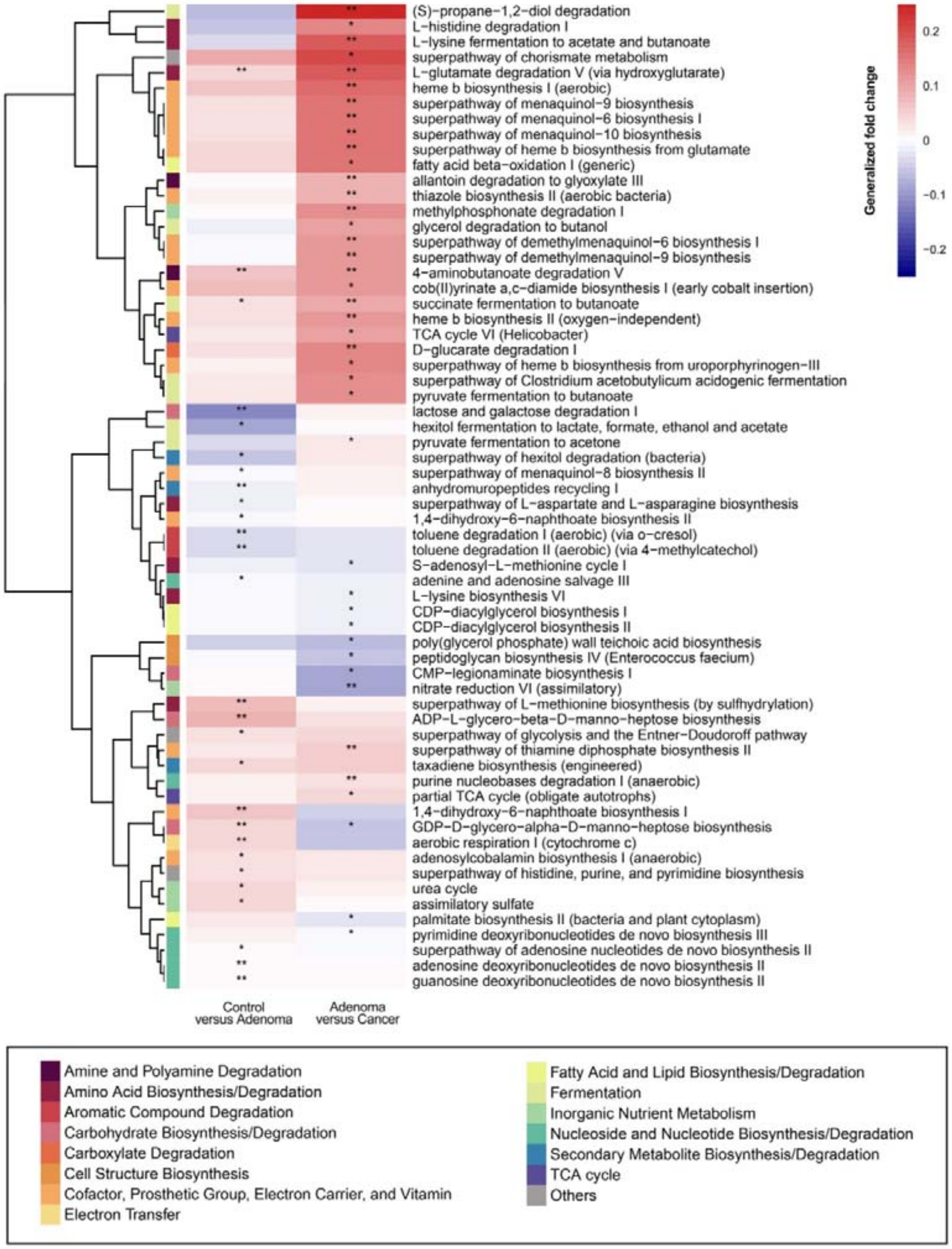

Fig. 4 Functional alterations in control, adenoma and cancer. The relative abundances of functional pathways were compared between adenoma and control or cancer. These pathways were differentially abundant ones, with *: $P < 0.05$, **: $P < 0.01$. Generalized fold change (see method) was indicated by a gradient color. The generalized fold change >0: enriched in later; generalized fold change <0: enriched in former.

Notably, the biosynthesis of ADP-heptose, a key metabolic intermediate in the biosynthesis of lipopolysaccharide (LPS) was significantly enriched in adenoma compared with control. It was associated with the activation of Nuclear factor-κB (NF-κB) and a strong pro-inflammatory response(26) which led to colorectal adenoma. The ASV was assigned as *Veillonella*, one of the biomarkers differentiating healthy controls from adenoma samples (Fig. 2a). It was highly ranked among all ASVs in the average contribution of the ADP-heptose. (Supplementary Table 11). There are five limiting steps catalyzed by genes of *hldE*, *rfaD*, *gmhA* and *gmhB* in the biosynthesis of ADP-heptose. These four genes were consistently enriched in adenoma compared with control (Supplementary Table 12). To further validate the results and explore its possibility of application in diagnosis, we analyzed the expression patterns of these genes based on qRT-PCR. As shown in Fig. 5a-d, the expressions of the *hldE* and *rfaD* gene were enriched in adenoma compared with control, in consistent with the picrust2 results, especially that the *hldE* gene was statistically significant.

Moreover, it is worth noting that menaquinone (Vitamin K2) biosynthesis was significantly enriched in cancer compared with adenoma. Especially, the MK-10 (one type of Vitamin K2) was mainly produced by *Bacteroides*, one of the biomarkers between adenoma and cancer (Fig. 2b), which reinforced the previous study that *Bacteroides* had high-level production of MK-10(27). The ASV assigned as *Bacteroides* ranked the 3rd and 4th in contribution to MK-10 biosynthesis in adenoma and cancer among all ASVs (Supplementary Table 13). Collectively, the production of Vitamin K2 by microbiota may serve as a response to compensate for induction of feedback inhibition in colorectal cancer cells(28). We found a significantly increased abundance of *menH*, *menF* and *menC* in CRC samples compared with that of control in pooled datasets by blocked Wilcoxon test (Supplementary Table 12). These results were further confirmed in adenoma and CRC by qRT-PCR on several patient samples (Fig. 5e-g), especially the *menH* and *menF* genes with statistical significances.

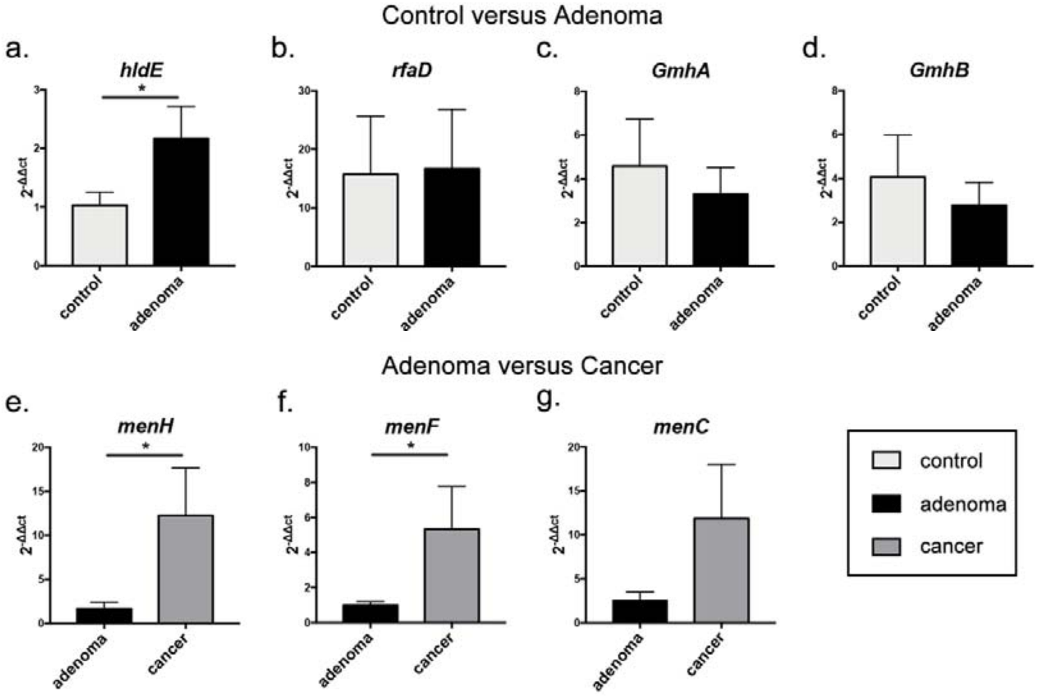

Fig. 5 Relative abundance of candidate genes is plotted against qRT-PCR quantification in gDNA extracted from stool samples of healthy controls, adenomas and CRC . The expression of (a) *hldE*; (b) *rfaD*; (c) *GmhA*; (d) *GmhB* were compared between control and adenoma groups, while the expression of (e) *menH*; (f) *menF*; (g) *menC* were compared between between adenoma and cancer groups. All results are presented as mean ± standard error. *: $P < 0.05$.

Discussion

This study comprehensively assessed the alterations of CRC-associated gut microbiome and the ability of microbial markers for the early detection of CRC. Thus, we first constructed machine learning classifiers. The best performing model achieves a high accuracy (AUC=0.80) with 11 important features to distinguish colorectal adenoma from non-tumor controls (Fig. 2c). Similarly, the AUC of the best model for detecting colorectal adenoma from CRC with 26 important features is 0.89 (Fig. 2d). Through study-to-study transfer validation and LODO validation across multiple datasets, the selected microbial markers could overcome technical and geographical discrepancies with the average AUC of 0.72 in the adenoma-control model (Fig. 3a) and 0.89 in the adenoma-cancer model (Fig. 3b), while previous researches revealed that the majority of differential microbial taxa differed in given case-control studies(17). Furthermore, the two additional independent cohorts strengthened and validated the extensibility of these makers (Supplementary Fig. 13a, b). The accuracy of classifiers with adenoma-specific markers is higher than that in previous WMS based studies(11, 14), probably due to more complete taxonomic profilings represented by ASVs. WMS data is well-recognized to possess the advantage of species- and even strain-level resolution. However, the current strategies for characterizing microbial community compositions with WMS are "closed annotation" that strongly rely on the known reference genome database(29-31), which is likely missing some species without known genomes or maker genes. It will thus result in biases in relative abundance estimation. Consistently, we built the cancer-control model with a panel of 34 important ASVs and achieved performance AUC of 0.93, whose accuracy is significantly higher than meta-analysis based WMS analysis(AUC=0.84)(11, 14). Importantly, the co-occurrence network analysis confirmed several biomarkers that are crucial in subnetworks, for example, *Ruminococcaceae UCG-005 sp.*, *[Clostridium] scindens*, *Blautia sp.*, etc. Furthermore, the performance of these features in diagnosing colorectal adenoma worsened when randomly adding samples from other microbiome-associated diseases, such as IBD and NAFLD(Supplementary Fig. 14), indicating that the panel of markers was adenoma-specific. Overall, all these validations point to the robustness of the classifiers and provide evidence that microbial classifier could serve as an effective non-invasive clinical indicator for colorectal adenoma.

Several other studies have reported that some fecal bacteria could serve as biomarkers for non-invasive diagnosis of colorectal cancer, such as *Fusobacterium nucleatum*, *Escherichia coli*, *Bacteroides fragilis*(8, 32-34). Unlike these existing studies, we aim to identify microbial-derived markers that could effectively diagnose adenoma (early stage of colorectal cancer), which represents a primary target for CRC screening at the early stage, as majority of CRC begins with the malignant transformation of benign polyps, the colorectal adenoma(2). Microbial communities altered in both colorectal adenoma and cancer during the progression of CRC, and the difference of microbiome alteration remains unclear. Notably, we found markers for distinguishing adenoma and cancers from healthy controls are not always the same (Supplementary Note 8). What's more, the combination of the important adenoma-specific features and FIT improved the classifier's accuracy (AUC=0.81) compared to microbial makers (AUC=0.78) or FIT (AUC=0.60) alone (Supplementary Fig. 11), indicating the non-invasive clinical screening tests could be used as complementary characteristics of gut microbiota for early screening of adenoma. Recently, a 16s rRNA analysis showed that microbiome dysbiosis in adjacent tissues could discriminate colorectal adenomas from healthy controls effectively(13), providing a new insight for following research of adenoma biomarkers.

The functional analysis sheds light on the convoluted underlying mechanisms and would greatly enhance our understanding and interpretation of CRC development (Supplementary Fig. 15). Among the differential pathways, we found the biosynthesis of ADP-heptose is significantly enriched in adenoma compared with control. ADP-heptose is a key metabolic intermediate in the biosynthesis of LPS, which is associated with the activation of NF-κB and then induces a strong pro-inflammatory response(35) in the initiation and progression of colorectal cancer, especially in adenoma(36). More importantly, the contribution decomposition analysis indicated that the adenoma-specific marker Veillonella parvula was highly ranked in the average contribution of the ADP-heptose among all ASVs (Supplementary Table 11). This suggests that the microbial markers contributed to the activation of pro-inflammatory pathways that ultimately led to the progression of colorectal adenoma. Notably, *hldE* was an important bifunctional protein involved in the biosynthesis of ADP-heptose, which catalyzes the nucleotide-activated heptose precursors used in the biosynthesis of LPS and in post-translational protein glycosylation(37). *HldE* was significantly enriched in adenoma compared with control according to computational finding and was further validated by qRT-PCR validation analysis(Fig. 5a). Since the *hldE* was reported to play an important role in bacterial virulence(37), it is promising to utilize it as an attractive target for therapeutic treatment of colorectal adenoma. Moreover, a series of Vitamin K2 biosynthesis is significantly different between adenoma and cancer. Especially, the MK-10 pathway increased in cancer compared with adenoma and the biomarker Bacteroides dorei was ranked as the third and fourth contributor to MK-10 biosynthesis in adenoma and cancer among all ASVs (Supplementary Table 13). Computational finding and qRT-PCR results demonstrated that the *menH* and *menF* gene were significantly increased in CRC compared with adenoma (Fig. 5e, f). Specifically, *menH* catalyzes the first step and plays distinct roles in the biosynthesis of MK-10 biosynthesis(38). The increased abundance of *menH* in CRC samples activated the synthesis of this pathway. Previous studies indicated that Vitamin K2 played a key role in antitumor effect via cell-cycle arrest, cell differentiation and cell apoptosis(28). Therefore, the increased production of Vitamin K2 may be a compensatory effect of the dysregulated microbiota to survive the tumor microenvironment, which also shows a potential novel CRC intervention strategy targeting Vitamin K2 biosynthesis bacteria. Though the main pathways differ between the control-adenoma and the adenoma-CRC stage, all these important pathways triggered by altered microbiome could offer promising perspectives and evidence for intervention and treatment in the CRC carcinogenesis (Supplementary Note 9).

Methods

Data collection

We collected data from studies in PubMed.gov that published 16S rRNA sequencing data on patients with CRC, adenomas and healthy controls. Only four studies with accessible metadata of every sample were included in this work. Raw sequencing data of these studies were downloaded from Sequence Read Archive (SRA) and European Nucleotide Archive (ENA) using identifiers: PRJNA389927 for Zeckular et al(12), PRJEB6070 for Zeller et al(20), PRJNA290926 for Baxter et al(22) and PRJNA362366 for Sze et al(21). Besides, two additional cohorts (Supplementary Table 8) were used as independent cohorts with accession numbers PRJNA534511(39) and PRJNA280026(40).

The collection of human data for real-time quantitative PCR (qRT-PCR) analysis was approved by the Review Board of School of Public Health, Shanghai Jiao Tong University School of Medicine. Patients were recruited for initial diagnosis and had never received any treatment before fecal sample collection. Patients with hereditary CRC syndromes, with a previous history of CRC were excluded from the study. Based on pathological section and colonoscopy results, recruited subjects were classified into three groups: (1) healthy subjects, namely controls: individuals with colonoscopy negative for tumor, adenoma or other diseases; (2) patients with adenoma: individuals with colorectal adenoma(s); and (3) patients with CRC: individuals with newly diagnosed CRC. A total of 94□subjects were initially recruited based on criteria sex, age, BMI and other confounding factors. Finally, 43 were remained: 30□patients with CRC, 6□adenomas and 7□controls. Stool was collected in fecal collection tubes and was stored at –80□°C. DNA was extracted from fecal samples using Stool Genomic DNA Kit (CW20925, CWBIO, China) following the manufacturer’s instructions. Relative gene expression by qRT-PCR and the patient characteristics for qRT-PCR were summarized in Supplementary Table 14.

Data Preprocessing

The 16S rRNA sequencing data were analyzed using Quantitative Insights Into Microbial Ecology (QIIME2 V.2018.11), a plugin-based platform for microbiome analysis(41). DADA2 software, wrapped in QIIME2, was used to filter out sequencing reads with quality score Q > 25 and denoise reads into amplicon sequence variants (ASVs) (i.e. 100% exact sequence match), resulting with feature tables and representative sequences. Taxonomy classification was assigned based on the naïve Bayes classifier using the classify-sklearn package(42) against the Silva-132-99 reference sequences. ASVs that couldn't be precisely annotated to species were reassigned to ones having the most similar sequences in the same genus (or family) using NCBI Blast. Subsequently, representative sequences were aligned using Fast Fourier Transform (MAFFT) in Multiple Alignment and a phylogenetic tree was generated with the Fast-Tree plugin. Then, the feature tables were converted to relative abundance tables. A set of ASVs that were confidently detectable in at least three studies and were present in at least 80% of samples were selected for further analysis.

Confounder analysis

We used ANOVA-like analysis(14) to quantify the effect of potential confounding factors and disease status. The total variance of a given ASV was compared to the variance explained by disease status (control, adenoma and cancer) and the variance by confounding factors (age, BMI, diabetes, Nonsteroidal anti-inflammatory drug (NSAID), platform, race, gender and study) akin to a linear model. Variance calculations were performed on ranks to account for non-Gaussian distribution of microbiome abundance data(14). Potential confounding factors with continuous values were transformed into discrete variables either as quartiles or in the case of BMI as groups of lean(>25), overweight (25-30), and obese(>30) based on conventional cutoffs.

Meta-analysis of differentially abundant ASVs

The significance of differential abundance was tested on a per ASV using the blocked Wilcoxon test implemented in the R (V.3.5.2) 'coin' package (*P* values < 0.05 were deemed as significant in all differential analysis). Confounder with high variance explanation was defined as a block to adjust the differential analysis. Significance was tested against a conditional null distribution derived from permutations of the observed data. Permutations were performed within 'study' to control variations in block size and composition(14). For further analysis, we evaluated a generalization of the (logarithmic) fold change for each ASV. This quantity is widely applied to genomic sequencing data such as RNA-seq and GRO-seq and further improved for better resolution of sparse microbiome profiles(43). The generalized fold change was calculated as the mean difference between predefined quantiles (ranging from 0.1 to 0.9 in increments of 0.1 in this study) of the logarithmic control and adenoma, and between adenoma and cancer distributions.

Model construction and features extraction

Following the differentially abundant ASVs analysis, we built Random Forest (RF) classifier models with stratified 10-fold cross-validation to distinguish adenoma from cancer or control. The features used for model building consisted of patient metadata as well as differential ASVs and alpha diversity indices. The alpha diversity indices consisted of Shannon Index, Simpson Index and Observed ASVs, while the patient metadata features consisted of age, gender and BMI. The RF classifier models were built with 501 estimator trees and each tree had 10% of the total features. Then an iterative feature elimination (IFE) step was used to optimize the performance of subsequent RF models. The top features from the top-performing model were selected as “important features” and the top microbial features as “biomarkers”. Finally, the AUC was used to evaluate the performance of the optimized models.

Co-occurrence and clustering analysis

To further analyze the co-occurrence of biomarkers, the relative abundances of biomarkers were discretized into binary values ‘positive’ or ‘negative. For each biomarker, the 90th percentile in control or adenoma was used as the threshold. A sample was labeled ‘positive’ when the relative abundance of ASV was above the defined threshold(14). Based on the binarized markers-by-sample matrix, biomarkers were then clustered using the Jaccard index. Associations between clusters and metadata were calculated by a Cochran–Mantel–Haenszel test with ‘study’ as blocking factors.

Model evaluation

To assess the generalizability of microbial-based adenoma classifiers across geographic and technical differences of metagenomic data generation and processing in multiple patient populations, both study-to-study transfer validation and leave-one-dataset-out (LODO) validation were performed. In study-to-study transfer validation, RF classifiers were trained in one single study and externally assessed on all other studies (off-diagonal cells in Fig. 3a-b). Meanwhile, we applied a nested cross-validation procedure on the training study to calculate within-study accuracy (diagonal cells in Fig. 3a, b). In LODO validation, data from one study was set as the testing set, while data from the remaining three studies were pooled as the training set. The input features of the validation classifiers were the important features identified from the IFE analysis.

To evaluate whether the selected important features would achieve the best performances in study-to-study transfer validation and LODO validation, we constructed RF models with 3 different sets of input features, including (1) all ASVs, (2) differential ASVs and (3) important features. Then we sought to identify if there was a minimal set of important features that could achieve higher accuracy. A few of the top-ranking important features were always included in the minimal set in prior. We used the same methods as the study-to-study transfer validation and LODO validation and then calculated the average AUC of each testing study as each point in Fig. 3c, d. Finally, we compared the predictive values in the testing set across RF models with different sets of input features.

Additional validation of independent studies and other diseases

As an external test, we used additional independent data to validate the performance of the selected important features to differentiate adenoma from cancer or control. The input features of RF models were the ASVs with the same taxonomy assignments as the selected important features as well as patient metadata (validation cohort2 without the patient metadata only used ASVs as input features).

To assess the specificity of the selected important features for colorectal adenoma, we investigated five non-CRC diseases. For each disease, we randomly drew 30 samples from the control group (excluding NAFLD and IBD diseases of which 15 samples were selected) as well as 30 samples from the cases, and added them to the validation cohort1 dataset in turns, labeled as controls. The random selection process was repeated ten times, and the validation AUC was computed accordingly.

Functional profile analysis

The functions of gut microbiome were inferred from 16S rRNA sequences with Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt2) as previously published(44). Functional metagenome profiles that have more than 20% samples with relative abundance < 1×10^{-5} and show up in less than three of the studies were removed. The differential analysis and generalized fold change calculations were performed on pathway profiles in the same way as on ASVs profiles (see data preprocessing method). Then, we evaluated the contribution of each ASV to overall differential pathways. The contribution was defined as the ratio of one ASV functional abundance to the total functional abundance of all ASVs in a given pathway.

Quantitative PCR validation

To quantify the abundance and expression of genes from two selected biosynthesis, the qRT-PCR analysis was performed on 14 healthy controls, 12 adenoma and 30 CRC samples. For these samples, the gDNA was extracted with the FecalGen DNA Kit (Cat# e9604) according to the manufacturer's instructions. We used the primes in the Supplementary Table 15 for candidate genes; standard primers F515 and R806 for 16S. To perform the qRT-PCR reaction, the final primer concentration was diluted to 0.5 μM including 5 ng of gDNA in a 20 μl final reaction volume with the SYBR Green qPCR Mix (Thermo Fisher Scientific). The used qRT-PCR program was as follows: pre-denaturation at 95 °C for 10 min; denaturation at 95 °C for 15 s for 40 cycles; annealing at 60°C for 60 s followed by melt curve analysis(14). The qRT-PCR analysis was to calculate $2^{-\Delta\Delta Ct}$ values between candidate genes and 16S Ct values. The significance of the comparison between adenoma and control or CRC samples was tested by Wilcoxon test (*P*<0.05).

Reference

1. Bray F, *et al.* (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 68(6):394-424. doi: 10.3322/caac.21492.
2. Wong SH & Yu J (2019) Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat Rev Gastroenterol Hepatol* 16(11):690-704. doi: 10.1038/s41575-019-0209-8.
3. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, & Brown ML (2011) Projections of the cost of cancer care in the United States: 2010-2020. *J Natl Cancer Inst* 103(2):117-128. doi: 10.1093/jnci/djq495.
4. Liang JQ, *et al.* (2020) A novel faecal Lachnoclostridium marker for the non-invasive diagnosis of colorectal adenoma and cancer. *Gut* 69(7):1248-1257. doi: 10.1136/gutjnl-2019-318532.

5. Ren ZG, *et al.* (2019) Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. *Gut* 68(6):1014-1023. doi: 10.1136/gutjnl-2017-315084.
6. Jiao N, *et al.* (2017) Suppressed Hepatic Bile Acid Signaling Despite Elevated Production of Primary and Secondary Bile Acids in Nafld. *Gastroenterology* 152(5):S1068-S1068. doi: Doi 10.1016/S0016-5085(17)33607-7.
7. Feng Q, *et al.* (2015) Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* 6:6528. doi: 10.1038/ncomms7528.
8. Yu J, *et al.* (2017) Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66(1):70-78. doi: 10.1136/gutjnl-2015-309800.
9. Coker OO, *et al.* (2019) Enteric fungal microbiota dysbiosis and ecological alterations in colorectal cancer. *Gut* 68(4):654-662. doi: 10.1136/gutjnl-2018-317178.
10. Nakatsu G, *et al.* (2018) Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology* 155(2):529-541 e525. doi: 10.1053/j.gastro.2018.04.018.
11. Thomas AM, *et al.* (2019) Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med* 25(4):667-678. doi: 10.1038/s41591-019-0405-7.
12. Zackular JP, Rogers MA, Ruffin MTt, & Schloss PD (2014) The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev Res (Phila)* 7(11):1112-1121. doi: 10.1158/1940-6207.CAPR-14-0129.
13. Mo Z, *et al.* (2020) Meta-analysis of 16S rRNA Microbial Data Identified Distinctive and Predictive Microbiota Dysbiosis in Colorectal Carcinoma Adjacent Tissue. *mSystems* 5(2):e00138-00120. doi: 10.1128/mSystems.00138-20.
14. Wirbel J, *et al.* (2019) Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* 25(4):679-689. doi: 10.1038/s41591-019-0406-6.
15. Yachida S, *et al.* (2019) Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med* 25(6):968-976. doi: 10.1038/s41591-019-0458-7.
16. Walters WA, Xu Z, & Knight R (2014) Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett* 588(22):4223-4233. doi: 10.1016/j.febslet.2014.09.039.
17. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, & Alm EJ (2017) Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* 8(1):1784. doi: 10.1038/s41467-017-01973-8.
18. Ternes D, *et al.* (2020) Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? *Trends in Microbiology* 28(5):401-423. doi: 10.1080/01621459.1972.10481279.
19. Rausch P, *et al.* (2019) Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome* 7(1):133. doi: 10.1186/s40168-019-0743-1.
20. Zeller G, *et al.* (2014) Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 10:766. doi: 10.15252/msb.20145645.

21. Sze MA, Baxter NT, Ruffin MTt, Rogers MAM, & Schloss PD (2017) Normalization of the microbiota in patients after treatment for colonic lesions. *Microbiome* 5(1):150. doi: 10.1186/s40168-017-0366-3.

22. Baxter NT, Koumpouras CC, Rogers MA, Ruffin MTt, & Schloss PD (2016) DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. *Microbiome* 4(1):59. doi: 10.1186/s40168-016-0205-y.

23. Wu J, Li Q, & Fu X (2019) Fusobacterium nucleatum Contributes to the Carcinogenesis of Colorectal Cancer by Inducing Inflammation and Suppressing Host Immunity. *Transl Oncol* 12(6):846-851. doi: 10.1016/j.tranon.2019.03.003.

24. Flemer B, *et al.* (2018) The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* 67(8):1454-1463. doi: 10.1136/gutjnl-2017-314814.

25. He Y, *et al.* (2018) Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med* 24(10):1532-1535. doi: 10.1038/s41591-018-0164-x.

26. Cong Y (2018) ALPK1: a pattern recognition receptor for bacterial ADP-heptose. *Precision Clinical Medicine* 1(2):57-59. doi: 10.1093/pcmedi/pby012.

27. Conly JM & Stein K (1992) The Production of Menaquinones (Vitamin-K2) by Intestinal Bacteria and Their Role in Maintaining Coagulation Homeostasis. *Progress in Food and Nutrition Science* 16(4):307-343.

28. Kawakita H, *et al.* (2009) Growth inhibitory effects of vitamin K2 on colon cancer cell lines via different types of cell death including autophagy and apoptosis. *Int J Mol Med* 23(6):709-716. doi: 10.3892/ijmm_00000184.

29. Segata N, *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* 9(8):811-814. doi: 10.1038/Nmeth.2066.

30. Milanese A, *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 10(1):1014. doi: 10.1038/s41467-019-08844-4.

31. Wood DE & Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15(3):R46. doi: 10.1186/gb-2014-15-3-r46.

32. Kostic AD, *et al.* (2012) Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* 22(2):292-298. doi: 10.1101/gr.126573.111.

33. Wu S, *et al.* (2009) A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat Med* 15(9):1016-1022. doi: 10.1038/nm.2015.

34. Gao R, *et al.* (2017) Dysbiosis signature of mycobiota in colon polyp and colorectal cancer. *Eur J Clin Microbiol Infect Dis* 36(12):2457-2468. doi: 10.1007/s10096-017-3085-6.

35. Zhou P, *et al.* (2018) Alpha-kinase 1 is a cytosolic innate immune receptor for bacterial ADP-heptose. *Nature* 561(7721):122-126. doi: 10.1038/s41586-018-0433-3.

36. Patel M, Horgan PG, McMillan DC, & Edwards J (2018) NF-kappaB pathways in the development and progression of colorectal cancer. *Transl Res* 197:43-56. doi: 10.1016/j.trsl.2018.02.002.

37. Hermansen GMM, *et al.* (2018) HldE is important for virulence phenotypes in enterotoxigenic Escherichia coli. *Frontiers in Cellular and Infection Microbiology* 8(253):1-13. doi: 10.3389/fcimb.2018.00253.

38. Dahm C, Muller R, Schulte G, Schmidt K, & Leistner E (1998) THE ROLE OF ISOCHORISMATE HYDROXYMUTASE GENES ENTC AND MENF IN ENTEROBACTIN AND MENAQUINONE BIOSYNTHESIS IN ESCHERICHIA COLI. *Biochimica et Biophysica Acta* 1425(2):377-386. doi: 10.1016/S0304-4165(98)00089-0.

39. Dadkhah E*, et al.* (2019) Gut microbiome identifies risk for colorectal polyps. *BMJ Open Gastroenterol* 6(1):e000297. doi: 10.1136/bmjgast-2019-000297.

40. Nakatsu G, *et al.* (2015) Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat Commun* 6:8727. doi: 10.1038/ncomms9727.

41. Bolyen E, *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37(8):852-857. doi: 10.1038/s41587-019-0209-9.

42. Pedregosa F, *et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825-2830.

43. Feng J*, et al.* (2012) GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* 28(21):2782-2788. doi: 10.1093/bioinformatics/bts515.

44. Douglas GM, *et al.* (2020) PICRUSt2: An improved and extensible approach for metagenome inference. *bioRxiv*:672295. doi: 10.1101/672295.

List of abbreviations:

ADP-heptose, ADP-L-glycero-beta-D-manno-heptose biosynthesis; ASVs, amplicon sequence variants; AUC, the area under the ROC curve; BMI, body mass index; CD, Crohn’s disease; CRC, Colorectal Cancer; ENA, European Nucleotide Archive; FIT, Faecal Immunochemical Test; FOBT, fecal occult blood test; IARC, International Agency for Research on Cancer; IBD, Inflammatory bowel disease; IBS, Irritable bowel syndrome; IFE, Iterative feature elimination; LODO, leave-one-dataset-out; LPS, lipopolysaccharide; MAFFT, Multiple Alignment using Fast Fourier Transform; MK-10, menaquinol-10 biosynthesis; NAFLD, Non-alcoholic fatty liver disease; NF-κB, Nuclear factor-κB; NSAID, Nonsteroidal anti-inflammatory drug; OTUs, operational taxonomic units; PICRUSt2, Phylogenetic Investigation of Communities by Reconstruction of Unobserved States 2; QIIME2: Quantitative Insights Into Microbial Ecology 2; qRT-PCR: real-time quantitative PCR; RF, Random Forest; ROC, Receiver operating characteristic; SRA, Sequence Read Archive; T2D, type 2 diabetes; UC, ulcerative colitis; WMS, whole metagenome shotgun sequencing.

Declarations

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Availability of data and materials:

Raw fastq files are available through the Sequence Read Archive(SRA) and European Nucleotide Archive (ENA), with project ID: PRJNA389927, PRJEB6070, PRJNA290926, PRJNA362366, PRJNA534511, PRJNA280026, PRJEB28350, PRJNA544721, PRJNA541332 and PRJNA82111.

Competing interests:

The authors declare that they have no competing interests.

Funding:

This work was supported by National Natural Science Foundation of China 81774152 (to RZ), 81770571 (to LZ), 31900129 (to NL), National Postdoctoral Program for Innovative Talents of China BX20190393 (to NJ), China Postdoctoral Science Foundation 2019M651568 (to DW), 2019M663252 (to NJ), Natural Science Foundation of Shanghai 16ZR1449800 (to RZ), the National Key Clinical Discipline of China, Funds from the University at Buffalo Community of Excellence in Genome, Environment and Microbiome (GEM) (to LZ), the Program for Young Eastern Scholar at Shanghai Institutions of Higher Learning QD2018016 (to NL), Shanghai Pujiang Program 18PJ1406600 (to NL), Innovative research team of high-level local universities in Shanghai, Medicine and Engineering Interdisciplinary Research Fund of Shanghai Jiao Tong Univesity YG2019QNB39 (to NL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author’s contributions:

LZ, NL, CT and RZ conceived and designed the project. Each author has contributed significantly to the submitted work. YW and NJ drafted the manuscript. RZ, YZ, DW, AW, SF, WG, YL, SC, XH, PL, CT, NL and LZ revised the manuscript. All authors read and approved the final manuscript.