




# Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models

Puzi Jiang,<sup>a</sup> Sicheng Wu,<sup>a</sup> Qibin Luo,<sup>b</sup> Xing-ming Zhao,<sup>c,d</sup>  Wei-Hua Chen<sup>a,e</sup>

<sup>a</sup>Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center for Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, China

<sup>b</sup>Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising, Germany

<sup>c</sup>Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China

<sup>d</sup>Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Ministry of Education, China

<sup>e</sup>College of Life Science, HeNan Normal University, Xinxiang, Henan, China

Puzi Jiang and Sicheng Wu contributed equally to this work. Author order was determined both alphabetically and in order of increasing seniority.

**ABSTRACT** Common intestinal diseases such as Crohn's disease (CD), ulcerative colitis (UC), and colorectal cancer (CRC) share clinical symptoms and altered gut microbes, necessitating cross-disease comparisons and the use of multidisease models. Here, we performed meta-analyses on 13 fecal metagenome data sets of the three diseases. We identified 87 species and 65 pathway markers that were consistently changed in multiple data sets of the same diseases. According to their overall trends, we grouped the disease-enriched marker species into disease-specific and disease-common clusters and revealed their distinct phylogenetic relationships; species in the CD-specific cluster were phylogenetically related, while those in the CRC-specific cluster were more distant. Strikingly, UC-specific species were phylogenetically closer to CRC, likely because UC patients have higher risk of CRC. Consistent with their phylogenetic relationships, marker species had similar within-cluster and different between-cluster metabolic preferences. A portion of marker species and pathways correlated with an indicator of leaky gut, suggesting a link between gut dysbiosis and human-derived contents. Marker species showed more coordinated changes and tighter inner-connections in cases than the controls, suggesting that the diseased gut may represent a stressed environment and pose stronger selection on gut microbes. With the marker species and pathways, we constructed four high-performance (including multidisease) models with an area under the receiver operating characteristic curve (AUROC) of 0.87 and true-positive rates up to 90%, and explained their putative clinical applications. We identified consistent microbial alterations in common intestinal diseases, revealed metabolic capacities and the relationships among marker bacteria in distinct states, and supported the feasibility of metagenome-derived multidisease diagnosis.

**IMPORTANCE** Gut microbes have been identified as potential markers in distinguishing patients from controls in colorectal cancer, ulcerative colitis, and Crohn's disease individually, whereas there lacks a systematic analysis to investigate the exclusive microbial shifts of these enteropathies with similar clinical symptoms. Our meta-analysis and cross-disease comparisons identified consistent microbial alterations in each enteropathy, revealed microbial ecosystems among marker bacteria in distinct states, and demonstrated the necessity and feasibility of metagenome-based multidisease classifications. To the best of our knowledge, this is the first study to construct multi-class models for these common intestinal diseases.

**Citation** Jiang P, Wu S, Luo Q, Zhao X-m, Chen W-H. 2021. Metagenomic analysis of common intestinal diseases reveals relationships among microbial signatures and powers multidisease diagnostic models. *mSystems* 6:e00112-21. <https://doi.org/10.1128/mSystems.00112-21>.

**Editor** Vanni Bucci, University of Massachusetts Medical School

**Copyright** © 2021 Jiang et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Xing-ming Zhao, [xmzhao@fudan.edu.cn](mailto:xmzhao@fudan.edu.cn), or Wei-Hua Chen, [weihuachen@hust.edu.cn](mailto:weihuachen@hust.edu.cn).

**Received** 1 February 2021

**Accepted** 8 April 2021

**Published** 4 May 2021

**KEYWORDS** gut dysbiosis, human microbiome, intestinal disease, machine learning-based disease classification, noninvasive disease diagnosis

In recent years, the incidences of several intestinal diseases, including inflammatory bowel disease (IBD) and colorectal cancer (CRC), have been increasing in developing countries while remaining high in major western countries, mostly due to industrial urbanization and Western life-styles (1–6). For example, IBD, comprising mainly Crohn's disease (CD) and ulcerative colitis (UC), has increased incidence in newly industrialized countries in Africa, Asia, and South America (7); populations previously considered "low risk," including Indian and Japanese populations, also witnessed significant increase in incidence (6). In addition, as the overall incidence of CRC remained high in major western countries, an alarming trend of increased risk has been observed in young adults (3, 4).

IBD and CRC share several symptoms, including rectal bleeding, abdominal pain, diarrhea, weight loss, and anemia; furthermore, CRC in young patients has similar ages of onset to IBD (<50 years) (8). In addition, patients with IBD are considered at high risk of developing colorectal cancer, due to the duration of inflammation and expansion of lesions. The accumulative risk of CRC in IBD patients is increasing over time (9, 10). The European Crohn's and Colitis Organisation (ECCO) and the American Gastroenterological Association (AGA) recommend that IBD patients need to strengthen CRC surveillance with colonoscopies. But long-period surveillance does not solve the problem because of deficiencies of regular colonoscopies in detecting dysplasia and other high-risk factors in elderly patients (11). It thus can be challenging to accurately separate these diseases in clinical practice, especially in their early stages and/or in younger patients; delay in diagnosis is common and can cause harm, as a recent study has pointed out (8).

Recent studies have suggested that IBD and CRC are linked with a complicated interplay of various components, involving genetics, environmental factors, gut microbiome, and immune system (12, 13). So far, a few hundred genes have been identified, whose mutation and/or dysregulation of expression were linked to increased risk of IBD and CRC (14–17). However, genetic factors can only explain a limited proportion of the disease incidence (18–20). Conversely, other factors are believed to be major contributors, especially gut microbes (13, 21–23). The latter, along with the metabolites and antibiotics produced through digesting nutrients from food, the host, and other microbes could play important roles in modulating host immunity and inflammation, maintaining gastrointestinal equilibrium and resisting alien invaders (24).

Fecal microbial dysbiosis in IBD and CRC has been observed, and subsequently utilized to generate predictive models for patient stratification and/or risk evaluation (23, 25–27). For example, IBD patients showed a reduction of taxa from the *Firmicutes* phylum and enrichment of pathogenic species (26, 28). Several studies showed the IBD subtypes CD and UC had distinctive gut microbiota and metabolic profiles, though results differ across studies (29). Gut microbes have been identified as potential markers in distinguishing patients from controls in IBD and CRC individually, as both the increase of pathogens and development of lesions in the gut contribute to the dysbiosis through affecting metabolic functions of bacteria (25, 30, 31).

However, binary models (i.e., models capable of distinguishing patients of a particular disease from controls) created for a single disease may lead to misdiagnosis, on account of some microbes commonly changed in diseases (32). Furthermore, most models, especially those available for IBD and/or its subtypes, were generated on data from a single population and may not perform well on other populations (26, 28). In addition, though limited by the use of 16S amplicon sequencing data with low resolution, a study across multiple diseases to search for disease-specific markers raised the issue of whether we could distinguish one gut illness from others using solely gut microbiome data (32).

In sum, it is necessary to perform cross-disease comparisons and generate multi-class models capable of distinguishing these common intestinal diseases, which can

have very similar symptoms and associate with consistent gut microbiome alterations. It is also necessary to perform meta-analysis to account for population-specific biases. Meta-analysis is a method combining diverse projects that helps us avoid biases from individual study (33); moreover, the latest surveys about CRC via meta-analysis suggest the necessity of collecting metagenomics data as much as possible to identify consistently altered microbes (34, 35).

In this research, we collected 13 metagenomic data sets for common intestinal diseases known to have strong links to gut microbiota, including three, three, and seven data sets for CD, UC, and CRC, respectively, and performed meta-analysis to (i) determine disease-specific and consistent microbial alterations; (ii) elucidate possible mechanisms underlying the altered species associated with different disease states; and (iii) generate high-performance multiclass models using taxonomic and metabolic profiles for easier and better clinical applications.

## RESULTS

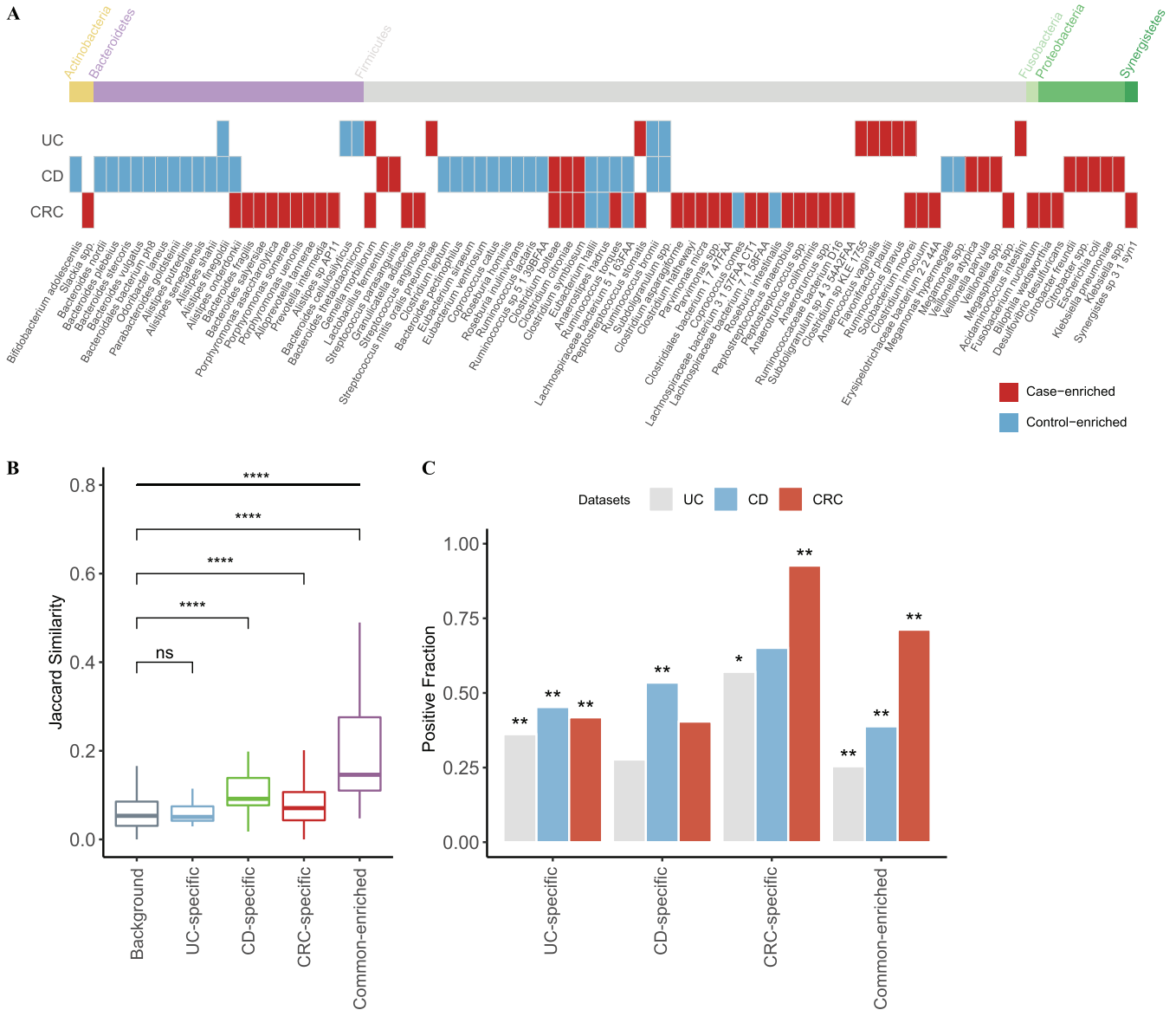
**Collection and annotation of 13 gut metagenomics data sets for common intestinal diseases.** To determine consistently altered gut microbial features in common intestinal diseases such as UC, CD, and CRC compared with controls, we conducted a systematic search in public databases (Fig. S1 in the supplemental material) and collected in total 13 metagenomic data sets, including three, three, and seven data sets for UC, CD, and CRC, respectively, totaling 763 cases and 632 controls (Table S1). MetaPhlan2 and HUMAnN2 were used to determine the taxonomic and functional profiles of all samples.

Taxonomic analysis revealed that the alpha diversity was not significantly changed in all but one CRC data set (PRJDB4176) compared with their respective controls; conversely, alpha diversity was decreased significantly in patients of most CD data sets, while it did not show consistent trends in UC patients (Table S1, Wilcoxon rank sum test,  $P$  value  $<0.05$ ). Interestingly, we found that the human DNA contents (HDCs), calculated as the percentage of sequencing reads mapped to the human genome, were significantly higher in patients in all diseases except one UC data set (PRJEB1220) (Table S1), consistent with our results (36) showing that HDC could be used as a marker for intestinal diseases; the increased level of HDCs are likely due to the high level of deciduous epithelial and/or blood cells found in stools of patients with IBD or CRC, resulting from gut injury and quickening cell cycles (22, 25, 30, 37).

**Disease-specific and shared taxonomic gut microbiome markers in CRC, UC, and CD.** We used MaAsLin2, a multivariable analysis tool, on the relative abundance of species to adjust the confounding factors, such as body mass index (BMI), gender, and age and identify differential species in each data set. We then performed meta-analysis on each disease to identify microbes that showed consistent trends in the same disease and referred to them as “marker species.” Consequently, we identified in total 14, 43, and 44 marker species in UC, CD, and CRC, respectively, among which 8 (57.1% out of 14), 32 (74.4%), and 31 (70%) were unique to the respective diseases. Out of a total of 87 marker species, 14 were found in at least two diseases and no one was common to all diseases (Fig. 1A).

For CRC, marker species were mostly disease-enriched, including *Fusobacterium nucleatum*, *Parvimonas micra*, *Gemella morbillorum*, and *Peptostreptococcus stomatis*, most of which were reported widely (34, 35). Interestingly, a significant proportion of the CRC-enriched marker species were significantly identified in a majority of CRC data sets, while most of CRC-depleted marker species were data set specific ( $P$  value  $<0.05$  identified by MaAsLin2, Fig. S2), which was in accordance with previous studies (34, 35).

Conversely, CD patients showed a depletion of control-enriched species, including *Roseburia inulinivorans*, *Roseburia hominis*, *Coprococcus catus*, and several members of the genera *Alistipes*, *Bacteroides*, and *Eubacterium*, which was also consistent with previous studies (26, 28). However, marker species in UC were a mix of both but mainly driven by disease-enriched ones (Fig. 1A), in contrast to a recent study that primarily



**FIG 1** Disease-specific and shared microbial markers showed distinct prevalence profiles in patients and controls. (A) Microbial markers and their trends (i.e., case- or control-enriched) in common intestinal diseases. Species significantly enriched in cases (or controls) of corresponding diseases in meta-analysis are shown ( $\text{fdr} < 0.05$  in meta-analysis, Benjamini-Hochberg FDR correction), with their phylum shown on top. Red indicates case-enriched species and blue indicates control-enriched ones. (B) Boxplot showing the inner Jaccard similarities of case-enriched microbes in all cases. The case-enriched microbes were clustered according to their trends in intestinal diseases (see the Materials and Methods). The term background indicates the similarities between members that did not belong to the same cluster. The four clusters were named according to their members; UC-specific, CD-specific, and CRC-specific clusters only contained the disease-specific markers, while the common-enriched cluster contained markers from at least two diseases ( $****, P < 0.0001$ ). (C) Barplot showing the fraction of cases that are “positive” for given clusters in a per-disease type. Here, positive samples for a given species are defined as those in which the species was found with higher relative abundance than 95% of all controls. The significant differences of the positive fraction between controls and cases for each cluster were assessed via Cochran-Mantel-Haenszel test with “data set” as the stratified factor; the asterisks indicate that marker species of a given cluster were significantly more prevalent in cases than in their corresponding controls ( $*$ ,  $P < 0.05$ ;  $**$ ,  $P < 0.01$ ).

showed a decrease in control-enriched species (28); the discrepancies are likely due to study-specific results (Fig. 1A). Moreover, most of the UC- and CD-marker species were identified as significant differential species in at least two data sets, unlike the CRC marker species ( $P$  value  $< 0.05$  identified by MaAsLin2, Fig. S2).

Among the shared markers, *Alistipes onderdonkii* and *Ruminococcus torques* showed conflicted trends between diseases; for example, they were both decreased in CD patients but increased in CRC patients. These results are in fact consistent with previous studies (38–41) and suggest that both overgrowth and loss of certain species represent the disturbance of the intestinal environment.

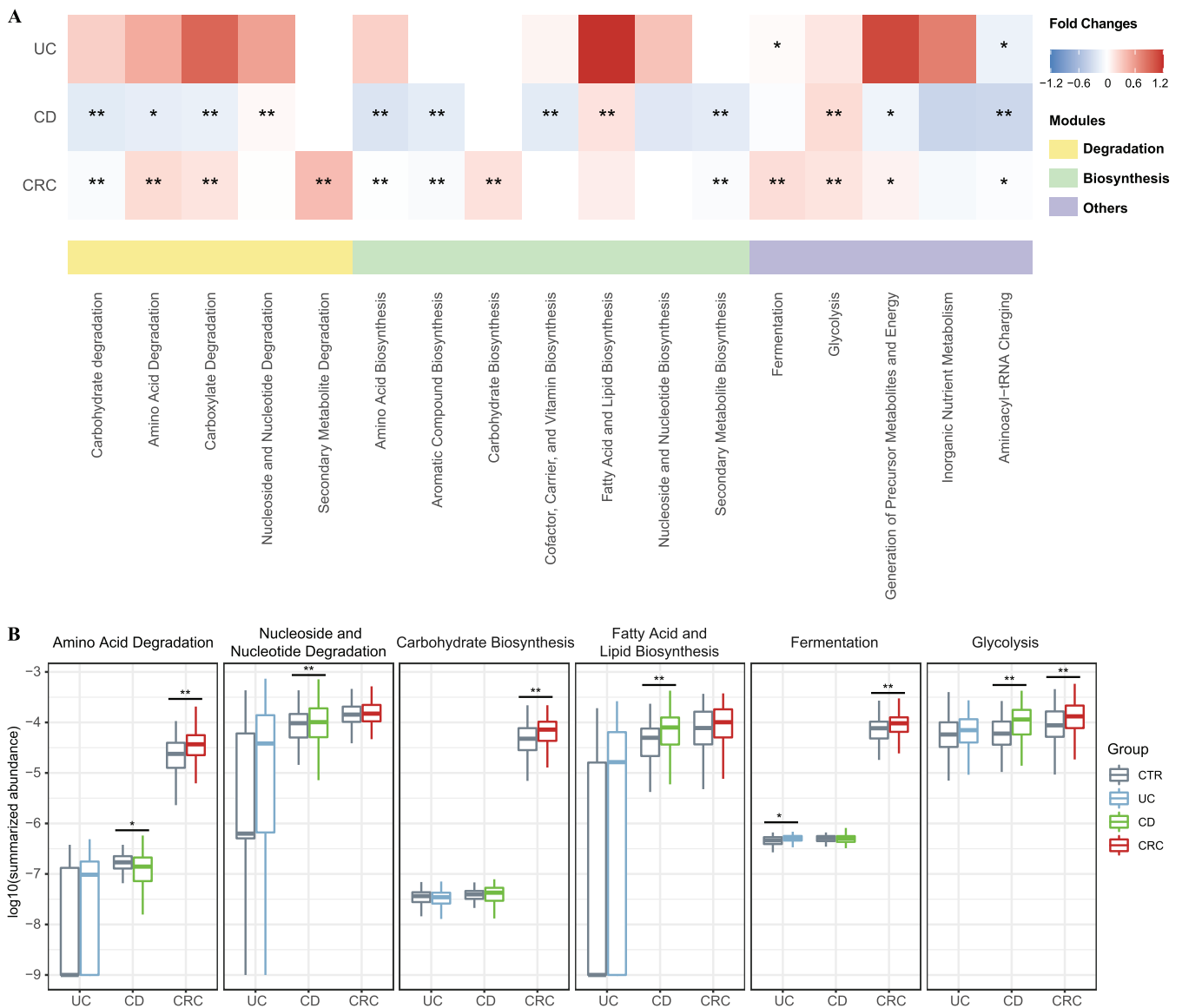
Disease-enriched species are often directly linked to pathogenesis and are direct targets for disease intervention. We thus first focused on these species and grouped them into disease-specific versus common according to their shifts in intestinal diseases. As shown in Fig. 1A, in total 6, 10, 31, and 6 markers were assigned to UC-specific, CD-specific, CRC-specific, and common-enriched groups, respectively. Their phylogenetic relationships based on the NCBI taxonomic tree revealed their distinct distributions (Fig. S3). First, CRC-specific markers, whose taxonomic levels fell across six phyla, showed more diverse phylogenetics than CD- and UC-specific markers. CD-specific markers were members of two phyla, namely, *Proteobacteria* and *Firmicutes*, including species within the *Veillonella* genus, *Enterobacteriaceae* family, and *Lactobacillales* order. The UC-specific markers consisted of species from *Firmicutes*. In addition, UC-specific species are phylogenetically closer to CRC-specific species, likely because UC patients have a little higher risk of CRC (9, 10). Together, we revealed phylogenetic patterns of the marker microbes that could only be revealed through cross-disease analysis.

We next checked if the disease-enriched species could also show distinct prevalence patterns in their respective diseases versus controls and/or in other diseases. To overcome the variances in species abundances, we defined a dynamic threshold for each species as its 95% quantile relative abundance of all control samples, and determined whether a species was present in a sample or absent (see the Materials and Methods). By doing so, we obtained a binarized matrix with each row representing a disease-enriched species and each column representing a patient. We calculated the Jaccard similarity to investigate the co-occurrence patterns of the clustered groups in all patients. As shown in Fig. 1B, we found the markers in the CD-specific, CRC-specific, and common groups showed significantly higher inner-similarity in patients, indicating these species were preferably able to coexist with the members in the same group; however, the UC-specific species did not display such a trend.

We then summed up the prevalence of the disease-enriched species of each sample. As expected, these disease-specific markers were significantly enriched in their respective diseases, together with the common microbes that were significantly enriched in all diseases (Fig. 1C). The UC-specific species, though, lacked a strong co-occurrence among themselves and had a remarkable prevalence in each disease data set. The results suggest that with the disease-specific clusters, it would be possible to stratify different diseases using microbial profiles, while the shared enriched species increased difficulties of classification.

**Disease-specific and shared functional markers in CRC, UC, and CD.** Using the same criteria, we identified in total 10, 37, and 39 marker pathways for UC, CD, and CRC, respectively (Fig. S4), among which 3 (30% out of 10), 18 (48.6%), and 25 (64.1%) were unique to these diseases, respectively. Most of the UC and CD marker pathways were control-enriched and associated with biosynthesis, consistent with previous results (28). For example, pathways for amino acid biosynthesis, such as L-methionine biosynthesis I and the aspartate superpathway, were depleted in CD patients, indicating that the microbiota was in favor of nutrient transport and uptake (29, 42). Conversely, the gut community in CRC showed distinct characteristics, with a decreased capacity for carbohydrate degradation and an increased capacity for amino acid degradation, which accorded with previous studies (34, 35).

To provide an overview on the changed functional capacities of gut microbes, we summarized the metabolic functions as the modules according to their superclasses in the MetaCyc database (43). We applied the differential abundance analysis in module level, and found the characteristic functional pattern of IBD and CRC (Fig. 2). As mentioned previously, the module of amino acid degradation was decreased in CD patients, while its trend behaved in the opposite way in CRC patients. In CD patients, the elevated module of nucleoside and nucleotide degradation, which was composed of the degradation of purine, would induce gut metabolic stress and involvement in inflammatory processes (44, 45). As essential pathways for energetic and biosynthetic



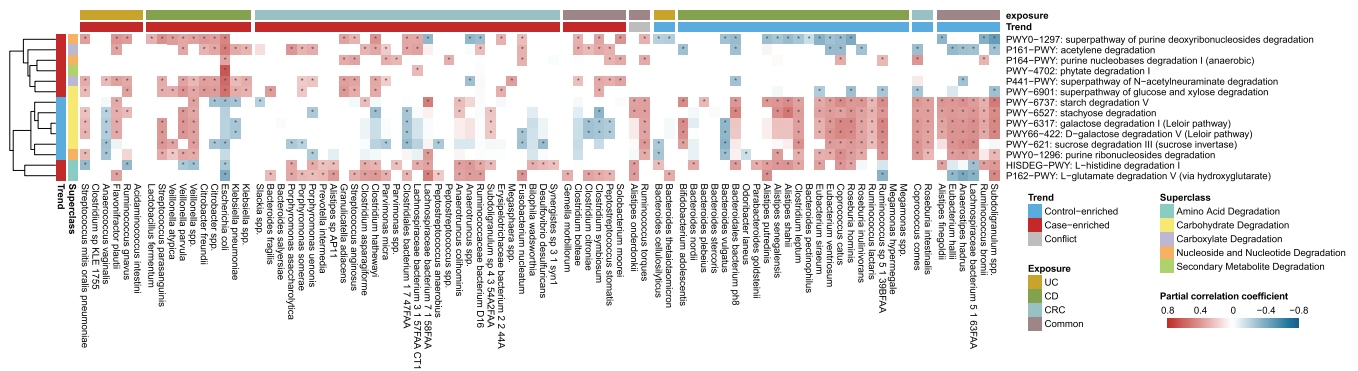
**FIG 2** Disease-specific pattern of metagenomics functional modules. (A) Functional modules were summed according to the category within the MetaCyc database. The differences between controls and cases in one specified disease were calculated as the generalized fold changes, and the significances were assessed using two-sided Wilcoxon rank sum tests and blocked with “data sets” (see the Materials and Methods). Red indicates case-enriched modules and blue indicates control-enriched modules. The asterisks indicate the modules were significantly different between cases and controls (\*,  $P < 0.05$ ; \*\*,  $P < 0.01$ ). (B) Boxplots showing the distribution of some representative modules in the three intestinal diseases. The asterisks on the bars were calculated as described above with the same significance values.

demands of cancer cells, the carbohydrate biosynthesis, fermentation, and glycolysis pathways were enhanced in CRC (46, 47). In addition, the subtypes of IBD (CD and UC) had distinct preferences at the module level, though there was a high degree of overlap among their associated pathways (Fig. S4).

Together, we identified consistently altered marker species and functional pathways in each of the intestinal diseases. A significant proportion of them were shared by two diseases, while the majority of them remained disease specific. Of note, UC was associated with the least number of disease markers and the least proportion of unique ones.

**Disease marker microbes underlie altered metabolic capacities, especially in degradation.** To check if the altered marker species could underlie the changes in metabolic capacities, we calculated partial Spearman’s rank correlations between marker species and metabolic pathways and performed meta-analysis to aggregate





**FIG 3** Marker microbes signified distinct degradation preferences. Shown here are the correlations of meta-analysis in relative abundances between the degradation pathways and the marker species (see the Materials and Methods). The pathways were clustered using the “mcquitty” algorithm, while the species were sorted by their related diseases and changing trends. The blocks in the heatmap show the overall coefficients from the meta-analysis. The red blocks indicate positive correlation and blue blocks indicate negative ones. The asterisk indicates that the adjusted *P* value of the overall coefficients in the meta-analysis is below 0.05. The similar plot for other pathways is shown in Fig. S5 in the supplemental material.

coefficients. Interestingly, we found that most of the disease-altered pathways showed statistically significant correlations with the marker species; more importantly, we were able to recapitulate the species clusters (including the control-enriched cluster identified in the previous sections) using their correlated metabolic capacities, especially in degradation (Fig. 3 and Fig. S5). For example, most of control-enriched species in CD and CRC, including members of the genera *Coprococcus*, *Roseburia*, *Ruminococcus*, and *Eubacterium*, were positively correlated with most carbohydrate degradation pathways, such as starch degradation V, stachyose degradation, galactose degradation I, and D-galactose degradation V (Fig. 3). These species are capable of fermenting general carbohydrates and producing butyrate, which has anti-inflammatory effects in the gut (48, 49). Additionally, a few disease-enriched pathways previously linked to CRC were also found to correlate with disease-enriched microbial markers. For example, *Lachnospiraceae* bacterium 7 1 58FAA had an evident link with L-glutamate degradation V, a CRC-specific pathway, via D-2 hydroxyglutarate that could drive epithelial-mesenchymal transition and induce CRC progression (50, 51). Similarly, some IBD-depleted species, such as *Alistipes shahii*, *Subdoligranulum* spp., and *Ruminococcus bromii*, had a negative association with the superpathway of purine deoxyribonucleoside degradation, a pathway used as the source of energy (52) (Fig. 3). Of note, the pathways relating to amino acid degradation were positively associated with most of the CRC-enriched bacteria, while negatively associated with IBD-enriched bacteria. Thus, clustered marker microbes could signify (at least in part) the changes in the overall metabolic capabilities in diseases and controls. In addition, these correlations between bacteria and microbial functions across studies and diseases revealed differences in metabolism among patients with different diseases, particularly between CRC and IBD.

The gut metabolic properties are known to be influenced by food and microbial activities. Recently, researchers also revealed that cells/metabolites derived from the human host, likely due to a compromised intestinal barrier (CIB), can also influence the growth of individual bacteria and the gut microbes as a whole (26, 29). CIB could lead to increased HDCs in the gut metagenomics. As expected, we found that HDCs were significantly elevated in cases of all data sets except PRJEB1220 (Table S1). Surprisingly, we found HDCs were also significantly correlated with some CRC-enriched species and half of control-enriched marker species in CD (Fig. S6). *Eubacterium ventriosum*, the control-enriched bacterial marker in CD data sets, was negatively correlated with HDCs in UC and CD ( $\rho = -0.32$ ,  $P$  value =  $8.16 \times 10^{-12}$  and  $\rho = -0.25$ ,  $P$  value =  $2.81 \times 10^{-6}$ , respectively, Spearman’s rank correlation); the correlation was not significant in CRC, most likely due to its low abundances. *E. ventriosum* was previously shown to negatively correlate with fundamental components of eukaryotic cell membranes (26). Only three control-enriched species in UC had associations with HDCs (Fig. S6A), while most

of the UC-depleted pathways correlated with HDCs (Fig. S6B), implying that the metabolic functions had a better response to the intestinal status.

Together, our results revealed correlated changes between marker species and metabolic pathways and suggested that both species and metabolic functions could be driven by the increased human-derived contents leaked into the gut due to CIB, consistent with our previous results (36).

**Marker species showed increased connectivity in diseases, presumably due to more stressed conditions.** Having shown that alterations in intestinal ecosystems could contribute to gut microbiota dysbiosis, we further explored the interrelationships among the marker species within each physical condition. As ecologically important patterns, coexistence relationships within a biological community could reflect interplays between organisms and ecological roles of individual members. Applying co-occurrence analyses to gut microbes could help us compare coexistence patterns from different intestinal states, identify key species important to human health, and provide an insight into the maintenance of gut microbial ecosystems (53, 54).

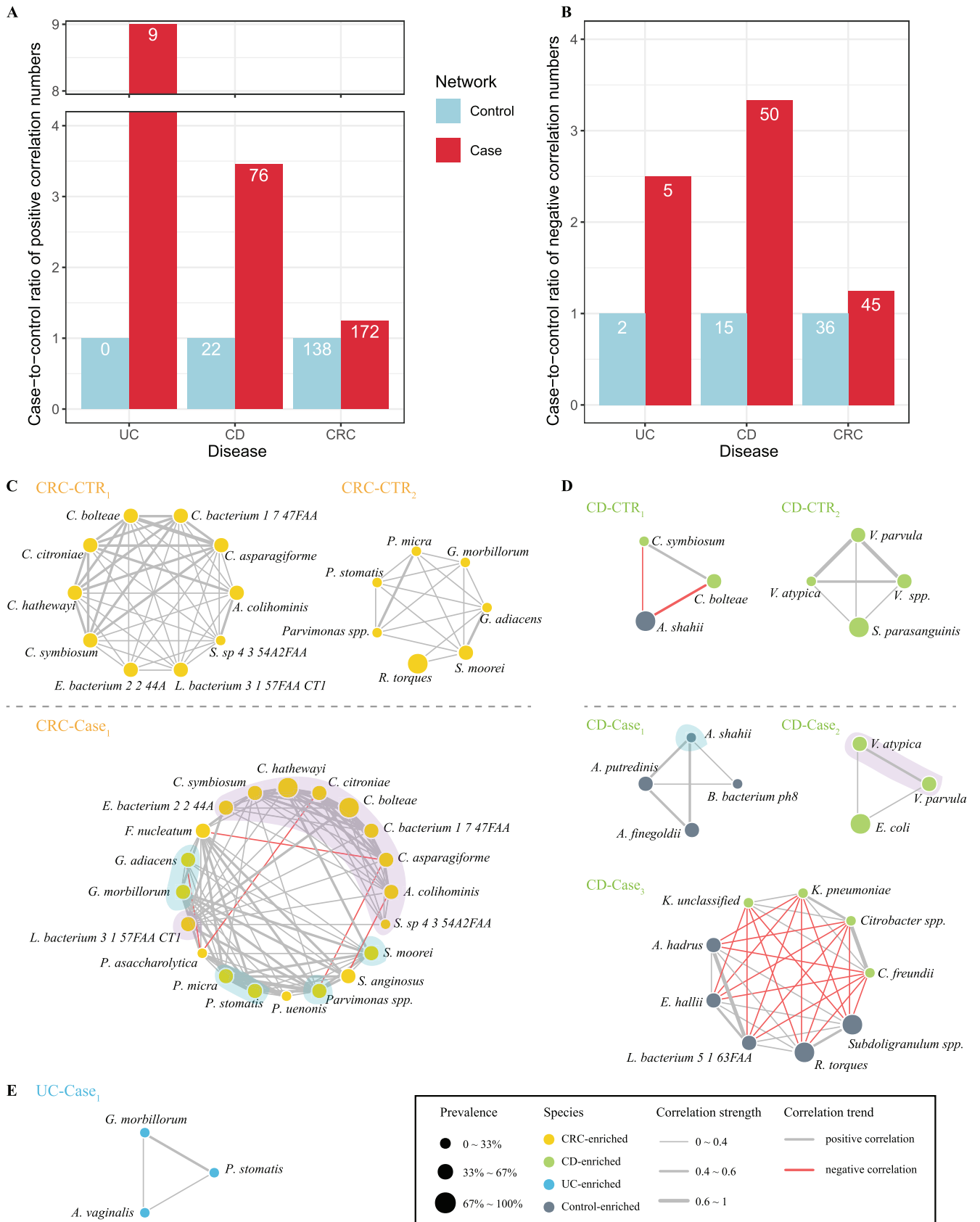
We thus constructed interspecies networks using the disease marker species separately for cases and controls for each disease based on pairwise correlations of the species abundances. We used SparCC, a correlation method for microbiome data, to calculate the correlation coefficients among species to perform meta-analysis. We found that species in the cases were connected more often than they were in the controls of respective disease data sets (Fig. 4). For example, we found 172 positive pairs and 45 negative pairs of correlated marker species ( $\text{fdr in meta-analysis} < 0.05$ ) in CRC, increased from 138 and 36 in the controls (Fig. 4A and B, Fig. S7A and B). Similarly, we found a greater number of positive and negative correlations among markers in cases than in controls in both CD and UC (Fig. 4A and B, Fig. S7C to F). These results were consistent with a previous study, which identified that the CRC patient networks contained more links among nodes than control networks, and the negative correlations declined when CRC patients underwent chemotherapy (55).

In view of the difficulties of comparing networks as a whole, we used the mcode implemented in Cytoscape to detect modules, which are regarded as the highly interconnected clusters in a network and often used to gain biological insights from networks. In CRC data sets, we found that the module from cases (named “CRC-Case1”) had more members and was interconnected tighter than those from controls (named “CRC-CTR1” and “CRC-CTR2”) (Fig. 4C). The species belonging to the same genus were associated more closely with each other, probably owing to their similar metabolic properties. Our network-derived modules could reveal previous known positive interactions. For example, *F. nucleatum*, a widely studied oral-associated anaerobe known to coaggregate with other anaerobes to form biofilm and involved in intestinal tumorigenesis (56, 57), showed positive correlations with *P. micra* in our case module (Fig. 4C) (58). Further, though it lacks experimental evidence for negative associations underlying CRC-enriched microbes, there have been numerous investigations about the competitive relationship among taxa during growth of biofilm (58–61). For example, *Porphyromonas gingivalis*, another known biofilm-forming partner of *F. nucleatum*, showed a negative correlation with *P. micra* (58). Thus, our results found novel relations between CRC-enriched microbes and remain to be confirmed in further experiments.

Similar to the CRC modules, modules from CD data sets also displayed tighter relationships within the species at same taxonomic level, including members of genera *Klebsiella*, *Veillonella*, and *Alistipes* (Fig. 4D). A previous study found that *Klebsiella* correlated positively with fecal calprotectin (FCP), an inflammatory marker for IBD, whereas *Ruminococcus* correlated negatively with FCP (54). In UC data sets, only the module from UC patient network (named “UC-Case1”) was recognized (Fig. 4E). The strong correlation between *G. morbillorum* and *P. stomatis* was shown in the UC module and CRC modules despite the sources of data, indicating the coaggregation between them.

We also identified potential hub species in the networks using eigenvector centrality scores (ECSs) and betweenness centrality scores (BCSs) (Fig. S8). ECS served as an





**FIG 4** Increased correlations of marker species in diseased conditions. (A and B) The statistics on the increases of positive correlated pairs (A) and negative correlated pairs (B) in the case network (red) compared with the control network (blue) in each of the three diseases. The y axis shows the case-to-control (Continued on next page)

assessment of node influence in a weighted network, measuring the importance of the given node considering not only its connections with others, but also the connections of its related nodes. BCS was used to evaluate the transmission capacity of species. In CD data sets, we found that *Alistipes putredinis*, *A. shahii*, and three CD-enriched *Veillonella* species were the pivotal species in controls, while in CD patients it was CD-enriched *Enterobacteriaceae*, *Citrobacter* spp., and *Klebsiella* spp. that took leading roles (Fig. S8C and D) and that could deliver virulence proteins into host cells to protect against the host immune system and infect mucosa, so as to thrive in gut (62–64). Nevertheless, health-related species were also the hub bacteria in patients with CD. These results support the view that we should be cautious in using antibiotics in CD therapy, as they may disrupt fragile connections among species existing in cases, and cause some bacteria failure of recovery (65). As expected, CRC-enriched microbes, such as species from the genus *Clostridium*, were at the center in both the control network and patient network of CRC data sets (Fig. S8A and B). Although the top nodes with high BCSs were CRC-enriched species in both the CRC control network and the CRC case network, their niches changed, suggesting that the CRC-enriched microbes were vital in the dynamic network. In UC data sets, nodes with the highest ECSs in the control network were UC-depleted species, while in the case network they were mainly UC-enriched species (Fig. S8E and F). We found the crucial nodes were also the members of the corresponding module, validating representativeness of the modules within a network.

Together, our interspecies network analysis revealed that marker species were more closely connected in diseased conditions; we speculated that due to oxidative stress and increased permeability of the intestinal barrier, the gut ecosystem under diseased states may represent more stressful conditions in which the growth of all microbes, especially the marker species, would be under stronger constraints and selection (12, 66). In addition, the hub species at the center positions and more connected with others are more likely to be targets for disease treatment.

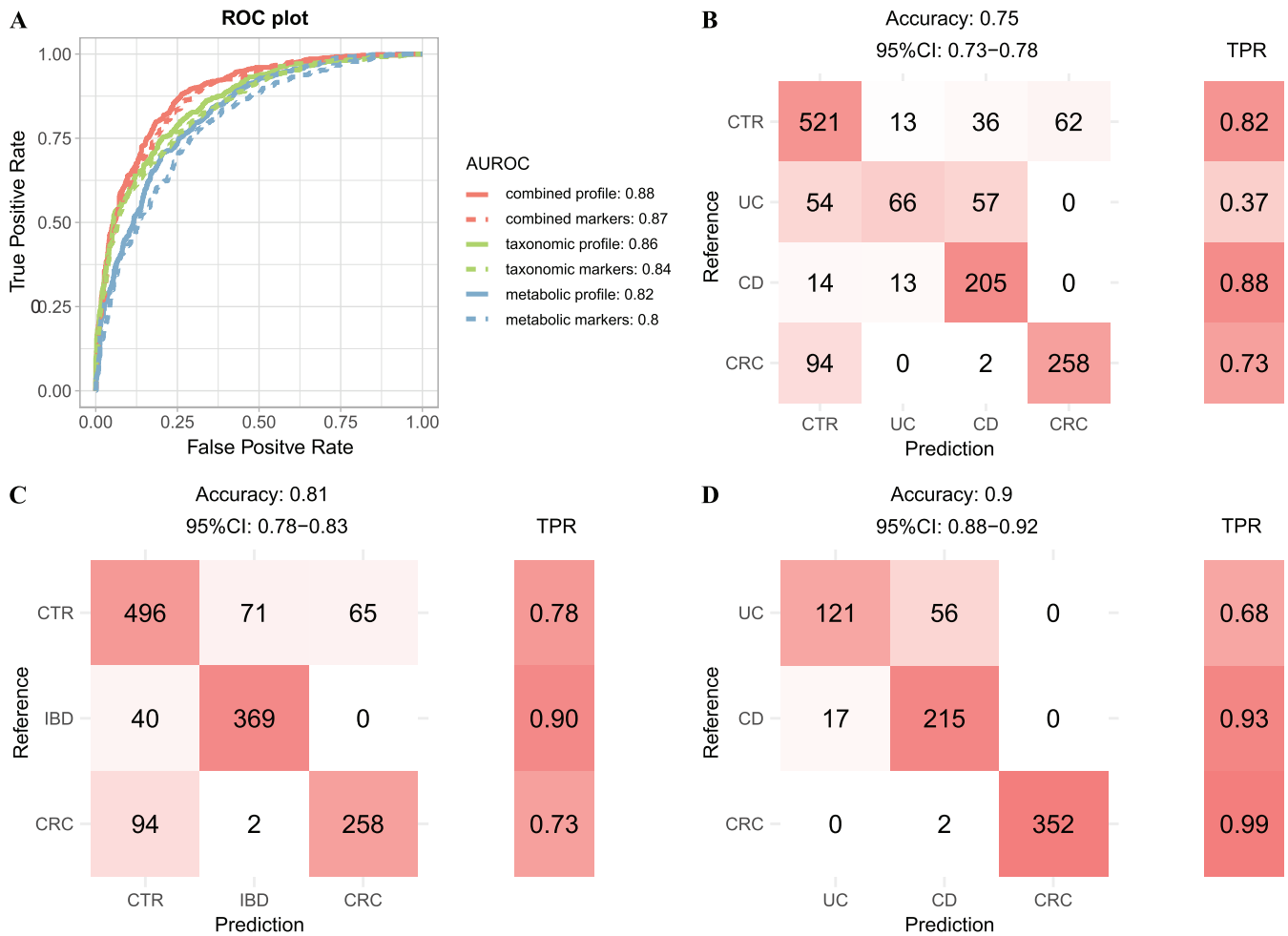
**Multiclass machine-learning models for disease stratification using gut microbial and metabolic markers.** We next built multiclass models capable of distinguishing multiple diseases using the microbial and pathway profiles, as well as those of the identified markers.

We first built four-class models with multiple types of features mentioned above using a 10-times and 10-fold cross-validation (see the Materials and Methods). A model based on combined taxonomic and metabolic profiles (all features) reached the highest accuracy, with a true-positive rate (TPR) of 0.8 (Fig. S9A and B). We achieved significantly better classification performance on UC and CD than a recent study, with TPRs of 0.81 and 0.91 for UC and CD samples, respectively, in this model (Fig. S9B), compared to 0.51 and 0.67 in reference 26 (see Fig. 6 reference 26), where the model was also trained on a combined profile. Models built on either taxonomic (TPR as 0.69) or functional profiles (0.74) showed decreased performance (Fig. S9A). The classifier based on combined markers also performed well with an overall accuracy of 0.75 (Fig. 5B). The classification errors were mostly contributed by UC samples, which associated with the least number of disease markers and the least proportion of unique ones (Fig. 1A and Fig. S4). Furthermore, UC shared a majority of its markers with CD, but not vice versa. Consequently, a significant proportion of UC samples were classified as CD, but only a few of the CD samples were misclassified as UC (Fig. 5B).

Since UC and CD are subtypes of IBD, we thus combined their samples into the IBD group and built three-class models (i.e., IBD, CRC, and controls, Fig. 5C and Fig. S9A). As shown in Fig. 5C, we achieved a much better classification rate with an overall TPR

#### FIG 4 Legend (Continued)

ratios of each disease, with the number of pairs in the control network being normalized to 1. The numbers in the bars indicate the actual numbers of correlated pairs. (C to E) Detected modules from correlation networks among marker species of per state in the corresponding disease data sets (see the Materials and Methods). Color of nodes means the alteration trends of species, and the sizes mean the prevalence of the bacteria in the overall health (or cases) of given data sets. Gray edges indicate positive relationship and red edges indicate negative. Thickness of edges indicates correlation strength.



**FIG 5** Random forest models for patient stratification using taxonomic and/or metabolic profiles. (A) Binary models for distinguishing cases from controls were constructed using various type of features as shown in the plot. The classification results were plotted as the AUROCs of the corresponding model. The AUROC was evaluated through 10-times 10-fold random forest cross-validation in all samples (see the Materials and Methods). The combined profile indicates the relative abundance profile of all taxonomic and metabolic features. The combined markers indicate the relative abundance of all taxonomic and metabolic markers. (B to D) Confusion matrix evaluation of the four-class model (B), three-class model (C), and cases model (D) based on combined marker features for distinguishing different physical conditions. The numbers in the boxes in the matrices on the left within each panel indicate how many patients with a given disease actually were categorized to that disease in the model. Color filling the cell means the relative size of the number in the corresponding row. The right part is the TPR for per-disease type. Total accuracy indicates the fraction of all correct predictions, and 95% CI is the confidence interval of accuracy.

of 0.81 using the combined markers; 90% of the IBD samples were classified correctly while the remaining 10% were misclassified as control, but none was classified as CRC.

We also built two additional models, including a “case-control” model to distinguish cases from controls (also referred to as “binary” model), and a “cases” model to assign cases to distinct diseased states (also referred to as “cases” model). The cases model was particularly important and clinically relevant due to the clinical overlap in presentation of these diseases, as well as the risk for IBD patients to eventually develop CRC. We evaluated the performance for the binary models on the 632 controls and 763 cases and found that all models performed well, with the area under receiver operating characteristic curves (AUROCs) ranging from 0.80 to 0.88. Notably, taxonomic-based models in general performed better than the metabolic-based models, as the model with the combined profile attained the highest accuracy (Fig. 5A). Surprisingly, models using only the marker species/pathways performed comparably to those that used all species/pathways, especially the combined model, suggesting that the much-shortened list of markers are of practical and clinical value. For the “cases” models, the classifier based on combined profile achieved an accuracy of 0.98, while the accuracy of

classifier based on combined markers only achieved 0.9. The metabolic-based model still performed better than the taxonomic-based model, indicating the functions of microbes reflected the gut status better than species distribution (Fig. 5D, Fig. S9A). We noticed significant differences among diseases in terms of TPRs in the “cases” models. In particular, we achieved high accuracies for CRC and CD (TPRs of 99% and 93%, respectively), compared with the relatively low TPR for UC (68%). The latter was likely due to the fact that UC shared most of its markers with CD and had only a few unique markers; consequently, most of the misclassified UC cases were predicted as CD (Fig. 1A and Fig. S4). Further, we analyzed the overlap between top features of the machine-learning classifiers and marker features and found there were 13, 22, 4, and 19 marker features in the top 30 features of the four-class, binary, three-class, and three-case classifiers built on the combined profiles, relatively (see the “Data availability” section).

To evaluate if the performances were being biased by a single data set, we applied leave-one-data set-out (LODO) analysis, which left one data set as the testing data and utilized the remained data sets to train the random forest models. The LODO models based on combined profile and combined markers (referred to as “four-class all” and “four-class dif”) to distinguish cases with the different disease and controls, achieved an average accuracy of 0.81 and 0.75 on training data, respectively, and an overall accuracy of 0.56 and 0.65 on testing data, respectively (Fig. S9C). The training results for each data set were similar, indicating there was no bias across data sets. Moreover, except for the models for classifying patients with different disease, the models trained on the combined markers performed better than the corresponding models trained on combined profiles.

These results suggest the classifiers based on combined markers could achieve similar accuracies with those based on combined profiles, indicating the clinical feasibility of the microbial markers. Besides, additional information other than fecal metagenomics, such as physiological, genetic, and clinical information on the human hosts are required to further improve the prediction accuracies.

## DISCUSSION

In this study, we collected fecal metagenomics data sets for three common intestinal diseases, namely, CRC, CD, and UC, totaling 11 projects, 13 data sets, 763 patients, and 632 controls. We selected these diseases because they all have strong associations with gut microbiota dysbiosis, share clinical presentations, and are pathogenically linked, i.e., both UC and CD patients are at high risk of developing CRC. We performed meta-analysis and identified in total 87 marker species and 65 marker pathways that were consistently changed (i.e., case-depleted or case-enriched) in the same disease. We grouped the marker species into disease-specific and disease-common clusters according to whether or not the member species were unique to a certain disease, and analyzed their distinct phylogenetic relationships; for example, CRC-specific species are more diverse phylogenetically than UC- and CD-specific markers. Strikingly, UC- and CRC-specific species are phylogenetically closer to each other than to those of CD (Fig. S3), in part due to the fact that UC patients are at higher risk of developing CRC (9, 10).

We then characterized the marker pathways. We first revealed that each disease formed their exclusive module profiles, that the CRC patients had an elevated trend in amino acid degradation while the CD patients behaved in an opposite way. We then showed that almost all marker pathways correlated significantly with marker species (Fig. 3 and Fig. S5); additionally, clustered marker species tended to correlate significantly with the same sets of pathways. These results were not unexpected since marker species that are closer phylogenetically tend to have similar metabolic capacities. We then noticed strong correlations between a significant proportion of marker species and HDCs. HDC has been shown to be significantly increased in many intestinal diseases, and could be used as an indicator for the extent of leaky gut caused by CIB (Table S1). The elevated HDCs in all data sets may signify significant changes in physio-metabolic properties

of the local gut environment due to leaked human-derived contents under diseased states. Our results thus suggested that human-derived contents due to CIB could have a stronger impact on gut microbiota than we have previously anticipated. Finally, by considering the gut microbiota as an ecosystem, we revealed that marker species showed increased connectivity in diseases compared with the respective controls, and control-enriched species together with pathogens played important roles in the ecological network of CD patients. Thus, we speculate that the diseased gut may represent a more stressful environment due to physio-metabolic changes, including oxidative stress and/or bleeding. If so, the inhabitant microbes are thus under stronger selection, and show either more cooperation (positive correlation) or competition (negative correlation). Our results support the view that we should be cautious in using antibiotics in therapies for CD patients.

Utilizing the identified marker species and pathways, we obtained four high-performance models for disease identification and patient stratification. The first “four-class” model could separate samples into controls or individual diseases (Fig. 5B), with an overall TPR of 0.75. UC has the lowest TPR (0.37, Fig. 5B) in this model; however, most of the wrongly classified samples went to CD, consistent with previous efforts (26) and with the fact that UC had very few unique markers and shared most of its markers with CD (Fig. 1A and Fig. S4). Regardless, it represents one of the best models that could classify IBD subtypes with TPR values of 0.81 and 0.91 separately in UC and CD samples, respectively, while in a previous study TPR values from the cross-validation model built on the abundance of metabolites and species only achieved 0.49 and 0.66 in UC and CD, respectively (26). We also built three additional models, including a “binary” model to distinguish cases from controls, a “three-class” model to consider the IBD as a whole and distinguish patients with cancer or inflammation from controls, and a “cases” model to assign cases to distinct diseased states. In our opinion, both are relevant in clinical applications. For example, the binary model, with an AUROC value of 0.87, can inform the subjects for further clinical inspections such as colonoscopy, while the “three-class” model, with an overall TPR of 0.81, can evaluate the patients for potential IBD and CRC risks. The “cases” model with a high accuracy of 0.9 was worth watching, due to the clinical overlap in symptoms of these intestinal diseases, as well as the risk for IBD patients to eventually develop CRC.

Taken together, our results demonstrated the necessity and feasibility of metagenome-based multidisease classifications. The few selected marker species and pathways had similar performances to all the taxonomic and metabolic features, and could be easily translated to clinical uses. Our meta-analysis methods and cross-disease comparisons improved our understanding of the differences and relationships among common intestinal diseases that could have similar clinical symptoms, and could be expanded to include more gastrointestinal disorders such as irritable bowel syndrome and colon polyps.

## MATERIALS AND METHODS

**Data collection and preprocessing.** We obtained in total 175 records by searching public metagenomic databases, including NCBI PubMed (67) and GMrepo (68), using key words such as metagenomics and relevant disease names (see Fig. S1 in the supplemental material for details). We aimed to collect metagenomic sequencing data with high resolution for better understanding the functions of microbes. After filtering out the duplicates, 16s rRNA sequencing data, and the metagenomics data without detailed metadata or not meeting minimum samples requirements, we selected in total 13 metagenomics data sets, including three, three, and seven data sets for CD, UC, and CRC, respectively. See Fig. S1 for the selection procedure and results, and see Table S1 for the 13 data sets.

Raw sequencing reads were retrieved from European Nucleotide Archive (ENA) (69) under the following identifiers: PRJEB6070 (23), PRJEB27928 (34), PRJEB12449 (70), PRJEB10878 (27), PRJEB7774 (40), PRJDB4176 (34), cohort 1 of PRJNA447983 (34), SRP057027 (25), PRJEB1220 (71), PRJNA400072 (26) and PRJNA389280 (72); sample metadata were also downloaded from ENA. For projects containing samples resulting from longitudinal surveys, i.e., participants were sampled multiple times over extended periods of time and/or during treatment/intervention, including SRP057027, PRJEB1220, and PRJNA389280, we selected the first time point from each participant to avoid false positives in the following analysis. In total, we obtained in 632 nondisease controls and 763 patients for the following meta-analysis, including 354, 177, and 232 samples of CRC, UC, and CD, respectively (Table S1).

**Taxonomic and functional profiling of metagenomics data.** To keep only the high-quality data, low-quality reads and adapters were first removed via Trimmomatic (version 0.35) using the TruSeq3 adapter files (TruSeq3-PE.fa for paired-end data and TruSeq3-SE.fa for single-end data) and a MINLEN

cutoff of 50 (73). The remaining “clean” reads were then mapped to the human reference genome (hg19) using bowtie2 (version 2.3.4.3) (74) with default settings to identify and remove human reads. The identified human reads were also used to compute HDCs for each sample as the percentage of mapped reads out of total clean reads, which have been shown to be a marker for intestinal barrier dysfunction and correlate with the marker species of several intestinal diseases (36). For samples that were sequenced multiple times (e.g., for the purpose of increasing sequencing depths), the resulting multiple sequencing files were merged before further analysis. The merged and clean nonhuman reads were then quantified in taxonomic and functional levels using MetaPhlan2 mapping to the mpa\_v20\_m200 database and HUMAnN2 mapping to the ChocoPhlan database and full UniRef90 database (75, 76).

To avoid the noise of low abundance, pathways with zero value in over 15% of samples within a data set were excluded. Species and pathways that did not meet a maximum relative abundance cutoff of  $1 \times 10^{-3}$  and  $1 \times 10^{-6}$  separately in at least 50% of data sets for a specified disease were removed. The abundance data were then loaded into R (ver 3.6.3 mainly; <https://www.r-project.org>) and analyzed.

**Controlling for confounding factors and identification of marker species and pathways.** Within-project confounding factors, i.e., those showing significant differences between phenotype groups in a data set, were first identified using a Wilcoxon rank sum test or chi-squared test on a per-data-set basis. Then, the identified confounding factors (see Table S1 for the results) in differential analysis were controlled for using MaAsLin2 package in R ver 4.0.0, a multivariable analysis tool to adjust the covariates and identify association effects of the species and pathways to disease in each data set. The species and pathways with raw *P* value below 0.05 in MaAsLin2 were considered differential species/pathways in the corresponding data set. Accounting for the heterogeneity between data sets, we performed meta-analysis to aggregate the association effects via MMUPHin package in R ver 4.0.0, and identified the final “marker” species and pathways. Here, an adjusted *P* value (fdr) of  $<0.05$  from meta-analysis was used as the cutoff for the markers.

**Clustering of disease-enriched species and their prevalence in the three diseases.** Consistently disease-enriched marker species (i.e., those that were marker species in at least two data sets of the same disease) were grouped into disease-specific or common to multiple diseases according to their association with the diseases (Fig. 1A). To observe the prevalence of the clusters in the overall patients versus a single disease (i.e., CRC, CD, or UC), their prevalence in the diseased samples were first calculated. For each selected marker species, its 95% percentile abundance in all controls was used as a cutoff to define its presence, where “1” indicated that the relative abundance in the sample was higher than the 95% quantile relative abundance of all control samples and “0” indicated absence. In this way, we obtained a binarized matrix with each row representing a disease-enriched marker microbe and each column representing a patient. The prevalence matrix from all patients was used to calculate the Jaccard distances among the species using the diversity function of the vegan package. We compared the inner Jaccard similarities among the clusters of disease-enriched marker species using a Wilcoxon rank sum test (for pairwise comparisons) and a Kruskal-Wallis rank sum test (for multigroup comparisons). The prevalence of each cluster between patients and controls in each disease was also compared using the Cochran-Mantel-Haenszel test with “data set” as the blocked object by the *cmh\_test* function of the coin package.

**Phylogenetic relationship of disease-enriched marker species.** To show the phylogenetic relationships among the disease-enriched marker species, a phylogenetic tree was generated based on their NCBI taxonomy using an online tool, phyloT (<https://phyloT.biobyte.de/>), setting internal nodes as collapsed and polytomy as no. The tree file then was visualized using Evolview ver3, a webserver for annotation and management of phylogenetic trees (77). The nodes were colored depending on their corresponding clusters as identified in the previous section. The last common ancestors (LCAs) were determined according to the NCBI taxonomy of the species in corresponding branch.

**Identification of HDC-correlated features.** For each data set, Spearman’s rank correlation was used to identify HDC-related microbial features (e.g., species and functions) using a *P* value cutoff of 0.05. Features that maintained a significant positive or negative relationship with HDCs in at least two data sets of a disease were identified as HDC-related features.

**Functional profile of metabolic modules.** According to the categories in the MetaCyc database, we grouped the microbial functions into their corresponding superclasses as metabolic modules (43). The expression of each metabolic module was summarized as the average logarithm relative abundance of its contained functions. Setting the quantiles from 0.1 to 0.9 and the increment as 0.1, we calculated the generalized fold changes of modules between the controls and cases, and performed the Wilcoxon rank sum test with the “data sets” as the blocked object to evaluate the differences.

**Microbial ecosystem analyses using species-species correlations.** To characterize the relationships among the marker species and the resulting interaction networks, SparCC, a sparse correlation method for compositional data (78) was used to identify correlations among marker species. SparCC was previously shown to be able to reduce the high false-positive rate by Spearman’s rank correlation in metagenomics data. The tool requires read counts as input, therefore we multiplied the relative abundances of the species to the number of reads mapping to mpa\_v20\_m200 database, and got the microbial counts of each sample. For each data set, species-species correlations were calculated for control and case samples separately. By setting both the iteration number and simulation as 100 and the threshold of correlation strength as 0.05, SparCC generated the correlation matrices of the real data and 100 simulated data sets. The pseudo *P* values were assessed as the proportion of simulated data sets with a correlation value at least as extreme as that calculated from the real data. After filtering correlations with *P* values of  $<0.05$ , we performed meta-analysis to aggregate correlation coefficients for each disease type via a random-effects model, which summarizes overall correlation based on Fisher’s *z*



transformation with metacor function (33, 79). The summarized correlations with adjusted  $P$  values of  $<0.05$  in meta-analysis were used to construct networks. The networks were analyzed in Cytoscape (80) to identify modules using mcode with default parameters. We then evaluated eigenvector centrality and betweenness centrality of networks using correlation strength as weight, and visualized networks with the igraph package in R. The size of the node indicated the prevalence of the bacteria in counterpart samples. Positive and negative correlation coefficients as strength of edges were painted gray and red separately.

**Correlating functional profiles with species.** To identify species underlying functional changes in the metagenomics data, correlations between relative abundance values of marker species and marker metabolic pathways in each data set were computed using partial Spearman's rank correlation to adjust the identified covariates. The relative abundances were log-transformed; to avoid Inf values, pseudo values of  $1e-06$  and  $1e-09$ , respectively, were added to the taxonomic and functional abundances before log-transformation. The resulting correlations with  $P$  values of  $<0.05$  were retained to perform meta-analysis and get the overall correlation coefficients with metacor function.

**Random forest classifiers and cross validation.** To check if the metagenomics data could be used to distinguish different diseases from each other and/or from healthy controls, the random forest function of the randomForest package were used to build several machine-learning classifiers. Samples were split into training and test data sets during the modeling. To prevent biases due to a one-time split, a 10-times and 10-fold cross-validation technique was used within the caret package. Thus, for each model, in total 100 models were created and the overall performance was the average of all the 100 models.

For the overall cross-validation, we pooled data sets into one, then applied logarithm transformation and standardization to the taxonomic and functional abundance profiles. The data were split into training set and testing set repeatedly 10 times. All models trained on the training set were applied to the corresponding testing set and the prediction scores were averaged. For binary classifiers, i.e., classifiers that attempt to classify samples into two distinct groups (diseased or control), the AUROC values were used to evaluate their performance. For the multiclass classifiers, i.e., classifiers that attempt to classify samples into multiple distinct groups such as CRC, UC, CD, and control, the detailed predicting results were shown as confusion matrixes and the performances as overall accuracies. We also built models based on the identified microbial markers to test if the performances of models were improved.

LODO analysis was performed to test if the cross-validation models were biased due to one specific data set. In short, all but one data set were pooled to build models with 10-fold and 10-times cross-validation as described earlier, and the resulting model was then applied to the left-out data set. To optimize each model after each split, we set the ranges for the number of trees and number of features to tune the hyperparameters with the mlr3 package. The whole process was repeated several times until every data set was used as the left-out data set in turn.

**Other statistical tests.** We calculated alpha diversity of each data set with the diversity function of the vegan package. A two-sided Wilcoxon rank sum test was used to compare two sets of numeric data; the wilcox\_test function implemented in the coin package was used to block with the factor "data set."

**Data availability.** The data sets generated and R codes during this study are available at <https://github.com/whchenlab/2019-puzi-multi-gut-disease-classifier>. Correspondence and requests for materials should be addressed to W.-H.C. and X.-m.Z.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**FIG S1**, PDF file, 0.1 MB.

**FIG S2**, PDF file, 0.3 MB.

**FIG S3**, PDF file, 0.3 MB.

**FIG S4**, PDF file, 0.3 MB.

**FIG S5**, PDF file, 0.04 MB.

**FIG S6**, PDF file, 0.4 MB.

**FIG S7**, PDF file, 0.5 MB.

**FIG S8**, PDF file, 0.2 MB.

**FIG S9**, PDF file, 0.2 MB.

**TABLE S1**, XLSX file, 0.01 MB.

## ACKNOWLEDGMENTS

We thank Na L Gao, Lei Liu, and Xinming Li for valuable discussion.

This work was partly supported by the National Key Research and Development Program of China (2019YFA0905600 to W.-H.C.), the National Natural Science Foundation of China (61932008, 61772368, and 61572363 to X.-m.Z.), the National Key Research and Development Program of China (2018YFC0910500 to X.-m.Z.), the Natural Science Foundation of Shanghai (17ZR1445600 to X.-m.Z.), a Shanghai Municipal Science and Technology Major Project (2018SHZDZX01 to X.-m.Z.), and ZJLab. The

fundes had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

W.-H.C. and X.-m.Z. designed the study; P.J. and S.W. collected and analyzed the data; Q.L. coordinated the data downloads and analysis; W.-H.C., and P.J. wrote the manuscript and all authors contributed to the writing and providing feedback. All authors read and approved the final version of the manuscript.

We declare no competing interests.

## REFERENCES

- Zhu J, Tan Z, Hollis-Hansen K, Zhang Y, Yu C, Li Y. 2017. Epidemiological trends in colorectal cancer in China: an ecological study. *Dig Dis Sci* 62:235–243. <https://doi.org/10.1007/s10620-016-4362-4>.
- Ye Y, Pang Z, Chen W, Ju S, Zhou C. 2015. The epidemiology and risk factors of inflammatory bowel disease. *Int J Clin Exp Med* 8:22529–22542.
- Vuik FE, Nieuwenburg SA, Bardou M, Lansdorp-Vogelaar I, Dinis-Ribeiro M, Bento MJ, Zadnik V, Pellise M, Esteban L, Kaminski MF, Suchanek S, Ngo O, Majek O, Leja M, Kuipers EJ, Spaander MC. 2019. Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. *Gut* <https://doi.org/10.1136/gutjnl-2018-317592>.
- Araghi M, Soerjomataram I, Bardot A, Ferlay J, Cabasag CJ, Morrison DS, De P, Tervonen H, Walsh PM, Burcher O, Engholm G, Jackson C, McClure C, Woods RR, Saint-Jacques N, Morgan E, Ransom D, Thursfield V, Moller B, Leonfellner S, Guren MG, Bray F, Arnold M. 2019. Changes in colorectal cancer incidence in seven high-income countries: a population-based study. *Lancet Gastroenterol Hepatol* 4:511–518. [https://doi.org/10.1016/S2468-1253\(19\)30147-5](https://doi.org/10.1016/S2468-1253(19)30147-5).
- M'Koma AE. 2013. Inflammatory bowel disease: an expanding global health problem. *Clin Med Insights Gastroenterol* 6:33–47. <https://doi.org/10.4137/CGast.S12731>.
- Ananthakrishnan AN. 2015. Epidemiology and risk factors for IBD. *Nat Rev Gastroenterol Hepatol* 12:205–217. <https://doi.org/10.1038/nrgastro.2015.34>.
- Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, Panaccione R, Ghosh S, Wu JCY, Chan FKL, Sung JY, Kaplan GG. 2018. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet* 390:2769–2778. [https://doi.org/10.1016/S0140-6736\(17\)32448-0](https://doi.org/10.1016/S0140-6736(17)32448-0).
- Stapley SA, Rubin GP, Alsina D, Shephard EA, Rutter MD, Hamilton WT. 2017. Clinical features of bowel disease in patients aged <50 years in primary care: a large case-control study. *Br J Gen Pract* 67:e336–e344. <https://doi.org/10.3399/bjgp17X690425>.
- Farraye FA, Odze RD, Eaden J, Itzkowitz SH. 2010. AGA technical review on the diagnosis and management of colorectal neoplasia in inflammatory bowel disease. *Gastroenterology* 138:746–774.e4. <https://doi.org/10.1053/j.gastro.2009.12.035>.
- Sebastian S, Hernández V, Myreliid P, Kariv R, Tsianos E, Toruner M, Marti-Gallostra M, Spinelli A, van der Meulen-de Jong AE, Yuksel ES, Gasche C, Ardizzone S, Danese S. 2014. Colorectal cancer in inflammatory bowel disease: results of the 3rd ECCO pathogenesis scientific workshop (I). *J Crohns Colitis* 8:5–18. <https://doi.org/10.1016/j.crohns.2013.04.008>.
- Abdalla M, Herfarth H. 2018. Rethinking colorectal cancer screening in IBD, is it time to revisit the guidelines? *J Crohns Colitis* 12:757–759. <https://doi.org/10.1093/ecco-jcc/jjy073>.
- Ni J, Wu GD, Albenberg L, Tomov VT. 2017. Gut microbiota and IBD: causation or correlation? *Nat Rev Gastroenterol Hepatol* 14:573–584. <https://doi.org/10.1038/nrgastro.2017.88>.
- O'Keefe SJD. 2016. Diet, microorganisms and their metabolites, and colon cancer. *Nat Rev Gastroenterol Hepatol* 13:691–706. <https://doi.org/10.1038/nrgastro.2016.165>.
- Khor B, Gardet A, Xavier RJ. 2011. Genetics and pathogenesis of inflammatory bowel disease. *Nature* 474:307–317. <https://doi.org/10.1038/nature10209>.
- Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, Abedian S, Cheon JH, Cho J, Daryani NE, Franke L, Fuyuno Y, Hart A, Juyal RC, Juyal G, Kim WH, Morris AP, Poustchi H, Newman WG, Midha V, Orchard TR, Vahedi H, Sood A, Sung JY, Malekzadeh R, Westra H-J, Yamazaki K, Yang S-K, Barrett JC, Franke A, Alizadeh BZ, Parkes M, Bk T, Daly MJ, Kubo M, Anderson CA, Weersma RK, International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium. 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics* 47:979–986. <https://doi.org/10.1038/ng.3359>.
- Ma X, Zhang B, Zheng W. 2014. Genetic variants associated with colorectal cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Gut* 63:326. <https://doi.org/10.1136/gutjnl-2012-304121>.
- Hu Z, Ding J, Ma Z, Sun R, Seoane JA, Scott Shaffer J, Suarez CJ, Berghoff AS, Cremolini C, Falcone A, Loupakis F, Birner P, Preusser M, Lenz H-J, Curtis C. 2019. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nature Genetics* 51:1113–1122. <https://doi.org/10.1038/s41588-019-0423-x>.
- Frank C, Sundquist J, Yu H, Hemminki A, Hemminki K. 2017. Concordant and discordant familial cancer: familial risks, proportions and population impact. *Int J Cancer* 140:1510–1516. <https://doi.org/10.1002/ijc.30583>.
- Foulkes WD. 2008. Inherited susceptibility to common cancers. *N Engl J Med* 359:2143–2153. <https://doi.org/10.1056/NEJMra0802968>.
- Knights D, Silverberg MS, Weersma RK, Gevers D, Dijkstra G, Huang H, Tyler AD, van Sommeren S, Imhann F, Stempak JM, Huang H, Vangay P, Al-Ghalith GA, Russell C, Sauk J, Knight J, Daly MJ, Huttenhower C, Xavier RJ. 2014. Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med* 6:107. <https://doi.org/10.1186/s13073-014-0107-1>.
- Lane ER, Zisman TL, Suskind DL. 2017. The microbiota in inflammatory bowel disease: current and therapeutic insights. *J Inflamm Res* 10:63–73. <https://doi.org/10.2147/JIR.S116088>.
- Dickinson BT, Kiesel J, Ahlquist DA, Grady WM. 2015. Molecular markers for colorectal cancer screening. *Gut* 64:1485–1494. <https://doi.org/10.1136/gutjnl-2014-308075>.
- Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Bohm J, Brunetti F, Habermann N, Herczog 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>.
- Rooks MG, Garrett WS. 2016. Gut microbiota, metabolites and host immunity. *Nat Rev Immunol* 16:341–352. <https://doi.org/10.1038/nri.2016.42>.
- Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, Bittinger K, Bailey A, Friedman ES, Hoffmann C, Albenberg L, Sinha R, Compher C, Gilroy E, Nessel L, Grant A, Chehoud C, Li H, Wu GD, Bushman FD. 2015. Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell Host Microbe* 18:489–500. <https://doi.org/10.1016/j.chom.2015.09.008>.
- Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, Sauk JS, Wilson RG, Stevens BW, Scott JM, Pierce K, Deik AA, Bullock K, Imhann F, Porter JA, Zhernakova A, Fu J, Weersma RK, Wijmenga C, Clish CB, Vlamakis H, Huttenhower C, Xavier RJ. 2019. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* 4:293–305. <https://doi.org/10.1038/s41564-018-0306-4>.
- Yu J, Feng Q, Wong SH, Zhang D, Liang Q, Qin Y, Tang L, Zhao H, Stenvang J, Li Y, Wang X, Xu X, Chen N, Wu WKK, Al-Aama J, Nielsen HJ, Kiilerich P, Jensen BAH, Yau TO, Lan Z, Jia H, Li J, Xiao L, Lam TYT, Ng SC, Cheng AS-L, Wong VW-S, Chan FKL, Xu X, Yang H, Madsen L, Datz C, Tilg H, Wang J, Brünner N, Kristiansen K, Arumugam M, Sung JY, Wang J. 2017. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66:70. <https://doi.org/10.1136/gutjnl-2015-309800>.
- Vich VA, Imhann F, Collij V, Jankipersadsing SA, Gurry T, Mujagic Z, Kurilshikov A, Bonder MJ, Jiang X, Tigchelaar EF, Dekens J, Peters V,

- Voskuil MD, Visschedijk MC, van Dullemen HM, Keszthelyi D, Swertz MA, Franke L, Alberts R, Festen EAM, Dijkstra G, Masclee AAM, Hofker MH, Xavier RJ, Alm EJ, Fu J, Wijmenga C, Jonkers D, Zhernakova A, Weersma RK. 2018. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci Transl Med* 10: eaap8914. <https://doi.org/10.1126/scitranslmed.aap8914>.
29. Schirmer M, Garner A, Vlamakis H, Xavier RJ. 2019. Microbial genes and pathways in inflammatory bowel disease. *Nature Rev Microbiology* 17:497–511. <https://doi.org/10.1038/s41579-019-0213-6>.
  30. Kim M, Ashida H, Ogawa M, Yoshikawa Y, Mimuro H, Sasakawa C. 2010. Bacterial interactions with the host epithelium. *Cell Host Microbe* 8:20–35. <https://doi.org/10.1016/j.chom.2010.06.006>.
  31. Zeng MY, Inohara N, Nuñez G. 2017. Mechanisms of inflammation-driven bacterial dysbiosis in the gut. *Mucosal Immunology* 10:18–26. <https://doi.org/10.1038/mi.2016.75>.
  32. 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* 8:1784. <https://doi.org/10.1038/s41467-017-01973-8>.
  33. Hunter J, Schmidt F. 1990. Dichotomization of continuous variables: the implications for meta-analysis. *J Applied Psychology* 75:334–349. <https://doi.org/10.1037/0021-9010.75.3.334>.
  34. 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>.
  35. 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. *Nature Medicine* 25:679–689. <https://doi.org/10.1038/s41591-019-0406-6>.
  36. Jiang P, Lai S, Wu S, Zhao X-M, Chen W-H. 2020. Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. *BMC Genomics* 21:348. <https://doi.org/10.1186/s12864-020-6749-z>.
  37. Vincent C, Mehrotra S, Loo VG, Dewar K, Manges AR. 2015. Excretion of host DNA in feces is associated with risk of *Clostridium difficile* infection. *J Immunol Res* 2015:246203. <https://doi.org/10.1155/2015/246203>.
  38. Alipour M, Zaidi D, Valcheva R, Jovel J, Martínez I, Sergi C, Walter J, Mason AL, Wong GK-S, Dieleman LA, Carroll MW, Huynh HQ, Wine E. 2016. Mucosal barrier depletion and loss of bacterial diversity are primary abnormalities in paediatric ulcerative colitis. *J Crohns Colitis* 10:462–471. <https://doi.org/10.1093/ecco-jcc/jjv223>.
  39. Takahashi K, Nishida A, Fujimoto T, Fujii M, Shioya M, Imaeda H, Inatomi O, Bamba S, Andoh A, Sugimoto M. 2016. Reduced abundance of butyrate-producing bacteria species in the fecal microbial community in Crohn's disease. *Digestion* 93:59–65. <https://doi.org/10.1159/000441768>.
  40. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, Su L, Li X, Li J, Xiao L, Huber-Schönauer U, Niederseer D, Xu X, Al-Aama JY, Yang H, Wang J, Kristiansen K, Arumugam M, Tilg H, Datz C, Wang J. 2015. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* 6:6528. <https://doi.org/10.1038/ncomms7528>.
  41. Zhang Y, Yu X, Yu E, Wang N, Cai Q, Shuai Q, Yan F, Jiang L, Wang H, Liu J, Chen Y, Li Z, Jiang Q. 2018. Changes in gut microbiota and plasma inflammatory factors across the stages of colorectal tumorigenesis: a case-control study. *BMC Microbiol* 18:92. <https://doi.org/10.1186/s12866-018-1232-6>.
  42. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 13:R79. <https://doi.org/10.1186/gb-2012-13-9-r79>.
  43. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD. 2020. The MetaCyc database of metabolic pathways and enzymes—a 2019 update. *Nucleic Acids Res* 48: D445–D453. <https://doi.org/10.1093/nar/gkz862>.
  44. Mars RAT, Yang Y, Ward T, Houtti M, Priya S, Lekatz HR, Tang X, Sun Z, Kalari KR, Korem T, Bhattarai Y, Zheng T, Bar N, Frost G, Johnson AJ, van Treuren W, Han S, Ordog T, Grover M, Sonnenburg J, D'Amato M, Camilleri M, Elinav E, Segal E, Blekhan R, Farrugia G, Swann JR, Knights D, Kashyap PC. 2020. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell* 183:1137–1140. <https://doi.org/10.1016/j.cell.2020.10.040>.
  45. Crittenden S, Cheyne A, Adams A, Forster T, Robb CT, Felton J, Ho GT, Ruckerl D, Rossi AG, Anderton SM, Ghazal P, Satsangi J, Howie SE, Yao C. 2018. Purine metabolism controls innate lymphoid cell function and protects against intestinal injury. *Immunol Cell Biol* 96:1049–1059. <https://doi.org/10.1111/imcb.12167>.
  46. La Vecchia S, Sebastián C. 2020. Metabolic pathways regulating colorectal cancer initiation and progression. *Semin Cell Dev Biol* 98:63–70. <https://doi.org/10.1016/j.semcdb.2019.05.018>.
  47. Olyphant K, Allen-Vercoe E. 2019. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome* 7:91. <https://doi.org/10.1186/s40168-019-0704-8>.
  48. Louis P, Hold GL, Flint HJ. 2014. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Rev Microbiol* 12:661–672. <https://doi.org/10.1038/nrmicro3344>.
  49. Louis P, Flint HJ. 2017. Formation of propionate and butyrate by the human colonic microbiota. *Environ Microbiol* 19:29–41. <https://doi.org/10.1111/1462-2920.13589>.
  50. Colvin H, Nishida N, Konno M, Haraguchi N, Takahashi H, Nishimura J, Hata T, Kawamoto K, Asai A, Tsunekuni K, Koseki J, Mizushima T, Satoh T, Doki Y, Mori M, Ishii H. 2016. Oncometabolite D-2-hydroxyglutarate directly induces epithelial-mesenchymal transition and is associated with distant metastasis in colorectal cancer. *Sci Rep* 6:36289. <https://doi.org/10.1038/srep36289>.
  51. Han J, Jackson D, Holm J, Turner K, Ashcraft P, Wang X, Cook B, Arning E, Genta RM, Venuprasad K, Souza RF, Sweetman L, Theiss AL. 2018. Elevated D-2-hydroxyglutarate during colitis drives progression to colorectal cancer. *Proc Natl Acad Sci U S A* 115:1057–1062. <https://doi.org/10.1073/pnas.1712625115>.
  52. Vich Vila A, Collij V, Sanna S, Sinha T, Imhann F, Bourgonje AR, Mujagic Z, Jonkers DMAE, Masclee AAM, Fu J, Kurilshikov A, Wijmenga C, Zhernakova A, Weersma RK. 2020. Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. *Nat Commun* 11:362. <https://doi.org/10.1038/s41467-019-14177-z>.
  53. Williams RJ, Howe A, Hofmockel KS. 2014. Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. *Front Microbiol* 5:358–358. <https://doi.org/10.3389/fmicb.2014.00358>.
  54. Yilmaz B, Juillerat P, Oyas O, Ramon C, Bravo FD, Franc Y, Fournier N, Michetti P, Mueller C, Geuking M, Pittet VEH, Maillard MH, Rogler G, Wiest R, Stelling J, Macpherson AJ, Swiss IBD Cohort Investigators. 2019. Microbial network disturbances in relapsing refractory Crohn's disease. *Nat Med* 25:701. <https://doi.org/10.1038/s41591-019-0411-9>.
  55. Cong J, Zhu J, Zhang C, Li T, Liu K, Liu D, Zhou N, Jiang M, Hou H, Zhang X. 2019. Chemotherapy alters the phylogenetic molecular ecological networks of intestinal microbial communities. *Front Microbiol* 10:1008. <https://doi.org/10.3389/fmicb.2019.01008>.
  56. Kostic Aleksandar D, Chun E, Robertson L, Glickman Jonathan N, Gallini Carey A, Michaud M, Clancy Thomas E, Chung Daniel C, Lochhead P, Hold Georgina L, El-Omar Emad M, Brenner D, Fuchs Charles S, Meyerson M, Garrett Wendy S. 2013. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 14:207–215. <https://doi.org/10.1016/j.chom.2013.07.007>.
  57. 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>.
  58. Horiuchi A, Kokubu E, Warita T, Ishihara K. 2020. Synergistic biofilm formation by *Parvimonas micra* and *Fusobacterium nucleatum*. *Anaerobe* 62:102100. <https://doi.org/10.1016/j.anaerobe.2019.102100>.
  59. Mark Welch JL, Ramírez-Puebla ST, Borisy GG. 2020. Oral microbiome geography: micron-scale habitat and niche. *Cell Host Microbe* 28:160–168. <https://doi.org/10.1016/j.chom.2020.07.009>.
  60. Palmer RJ, Shah N, Valm A, Paster B, Dewhirst F, Inui T, Cisar JO. 2017. Interbacterial adhesion networks within early oral biofilms of single human hosts. *Appl Environ Microbiol* 83:e00407-17. <https://doi.org/10.1128/AEM.00407-17>.

61. Bowen WH, Burne RA, Wu H, Koo H. 2018. Oral biofilms: pathogens, matrix, and polymicrobial interactions in microenvironments. *Trends Microbiol* 26:229–242. <https://doi.org/10.1016/j.tim.2017.09.008>.
62. Bengoechea JA, Sa Pessoa J. 2019. Klebsiella pneumoniae infection biology: living to counteract host defences. *FEMS Microbiol Rev* 43:123–144. <https://doi.org/10.1093/femsre/fuy043>.
63. Dinakaran V, Mandape SN, Shuba K, Pratap S, Sakhare SS, Tabatabai MA, Smoot DT, Farmer-Dixon CM, Kesavalu LN, Adunyah SE, Southerland JH, Gangula PR. 2019. Identification of specific oral and gut pathogens in full thickness colon of colitis patients: implications for colon motility. *Front Microbiol* 9:3220. <https://doi.org/10.3389/fmicb.2018.03220>.
64. Sontag RL, Nakayasu ES, Brown RN, Niemann GS, Sydor MA, Sanchez O, Ansong C, Lu S-Y, Choi H, Valleau D, Weitz KK, Savchenko A, Cambronne ED, Adkins JN. 2016. Identification of novel host interactors of effectors secreted by Salmonella and Citrobacter. *mSystems* 11:e00032–15. <https://doi.org/10.1128/mSystems.00032-15>.
65. Ananthkrishnan AN, Bernstein CN, Iliopoulos D, Macpherson A, Neurath MF, Ali RAR, Vavricka SR, Fiocchi C. 2018. Environmental triggers in IBD: a review of progress and evidence. *Nat Rev Gastroenterol Hepatol* 15:39–49. <https://doi.org/10.1038/nrgastro.2017.136>.
66. Dam B, Misra A, Banerjee S. 2019. Role of gut microbiota in combating oxidative stress, p 43–82. *In* Chakraborti S, Chakraborti T, Chattopadhyay D, Shaha C (ed), *Oxidative stress in microbial diseases*. Springer Singapore, Singapore. [https://doi.org/10.1007/978-981-13-8763-0\\_4](https://doi.org/10.1007/978-981-13-8763-0_4).
67. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A, Kitts PA, Kuznetsov A, Lathrop S, Lu Z, McGarvey K, Madden TL, Murphy TD, O'Leary N, Phan L, Schneider VA, Thibaud-Nissen F, Trawick BW, Pruitt KD, Ostell J. 2020. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 48:D9–D16. <https://doi.org/10.1093/nar/gkz899>.
68. Wu S, Sun C, Li Y, Wang T, Jia L, Lai S, Yang Y, Luo P, Dai D, Yang YQ, Luo Q, Gao NL, Ning K, He LJ, Zhao XM, Chen WH. 2020. GMrepo: a database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Res* 48:D545–D553. <https://doi.org/10.1093/nar/gkz764>.
69. Harrison PW, Ahamed A, Aslam R, Alako BTF, Burgin J, Buso N, Courtot M, Fan J, Gupta D, Haseeb M, Holt S, Ibrahim T, Ivanov E, Jayatilaka S, Balavenkataraman Kadhirvelu V, Kumar M, Lopez R, Kay S, Leinonen R, Liu X, O'Cathail C, Pakseresht A, Park Y, Pesant S, Rahman N, Rajan J, Sokolov A, Vijayaraja S, Waheed Z, Zyoud A, Burdett T, Cochrane G. 2021. The European Nucleotide Archive in 2020. *Nucleic Acids Res* 49:D82–D85. <https://doi.org/10.1093/nar/gkaa1028>.
70. Vogtman E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, Goedert JJ, Shi J, Bork P, Sinha R. 2016. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS One* 11:e0155362. <https://doi.org/10.1371/journal.pone.0155362>.
71. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jorgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clement K, Dore J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker JD, Raes J, Hansen T, Bork P, Wang J, Ehrlich SD, Pedersen O, MetaHIT Consortium. 2013. Richness of human gut microbiome correlates with metabolic markers. *Nature* 500:541–546. <https://doi.org/10.1038/nature12506>.
72. Schirmer M, Franzosa EA, Lloyd-Price J, McIver LJ, Schwager R, Poon TW, Ananthkrishnan AN, Andrews E, Barron G, Lake K, Prasad M, Sauk J, Stevens B, Wilson RG, Braun J, Denson LA, Kugathasan S, McGovern DPB, Vlamakis H, Xavier RJ, Huttenhower C. 2018. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol* 3:337–346. <https://doi.org/10.1038/s41564-017-0089-z>.
73. 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>.
74. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>.
75. 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>.
76. Franzosa EA, McIver LJ, Rahnava G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* 15:962. <https://doi.org/10.1038/s41592-018-0176-y>.
77. Subramanian B, Gao S, Lercher MJ, Hu S, Chen WH. 2019. Evolvview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res* 47:W270–W275. <https://doi.org/10.1093/nar/gkz357>.
78. 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>.
79. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods* 1:97–111. <https://doi.org/10.1002/jrsm.12>.
80. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. 2019. Cytoscape Automation: empowering workflow-based network analysis. *Genome Biol* 20:185. <https://doi.org/10.1186/s13059-019-1758-4>.