Metagenomic analysis of common intestinal diseases reveals relationships among microbial signatures and powers multi-disease diagnostic models

Puzi Jiang [1], Sicheng Wu [1], Xing-ming Zhao [2,3 §], Wei-Hua Chen [1,4 §]

Affiliations:

[1] Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, 430074 Wuhan, Hubei, China

[2] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China

[3] Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Ministry of Education, China

[4] College of Life Science, HeNan Normal University, 453007 Xinxiang, Henan, China

[§] To whom correspondence should be addressed

Emails:

Puzi Jiang: puzjiang@hust.edu.cn

Sicheng Wu: wu_sicheng@hust.edu.cn

Xing-Ming Zhao: xmzhao@fudan.edu.cn

Wei-Hua Chen: weihuachen@hust.edu.cn

Abstract

Intestinal diseases can share clinical symptoms and gut microbiome alterations, necessitating cross-disease comparisons and the use of multi-disease models. Here, we performed meta-analyses on fecal metagenomes of Crohn's disease (CD), ulcerative colitis (UC) and colorectal cancer (CRC), and identified 65 marker species and 74 marker pathways that were significantly altered in >=2 studies of the same disease. We grouped marker species by their co-occurrences into disease-specific and mixed clusters according to whether all the cluster members belong to the same disease, and revealed their distinct phylogenetic relationships: species in disease-specific clusters are phylogenetically related, while those in mixed clusters are more distant; strikingly, UC- and CRC-specific species are phylogenetically closer than to CD-specific ones, likely because UC-patients have higher risk of CRC. Consistent to their phylogenetic relationships, marker species had similar within-cluster and different between-cluster metabolic preferences. In CD and CRC, most marker species correlated with an indicator of leaky gut, suggesting a link between gut dysbiosis and human derived contents. In fact, our results suggested that due to physio-metabolic alterations, the diseased gut may represent a stressed environment and pose stronger selection to gut microbes; consequently, marker species in diseases showed more coordinated changes than the controls. With the marker species and pathways, we constructed three high-performance (including multi-disease) models with AUROC of 0.88 and true positive rates up to 89%,

and explained their putative clinical applications. Together, our results improved our understanding on the relationships among common intestinal diseases, and strongly support the feasibility of metagenome-based multi-disease diagnosis.

Background

In recent years, the incidences of several intestinal diseases including inflammatory bowel disease (IBD) and colorectal cancer (CRC) have been increasing in developmental countries while remained 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 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, diarrhoea, weight loss, and anaemia; 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 duration of inflammation and expansion of lesions. The accumulative risk of CRC in IBD patients is increasing decade by decade [9]. The European Crohn's and Colitis Organisation (ECCO) and the American Gastroenterological Association (AGA) recommend the IBD patients need to strengthen CRC surveillance with colonoscopies. But long-period surveillance doesn't solve the problem because of deficiencies of regular colonoscopies in detecting dysplasia and other high risk factors in the elderly patients [10]. 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

research has pointed out [8].

Recent studies suggested that IBD and CRC are linked with complicated interplay of various components, involving genetics, environmental factors, gut microbiome and immune system [11, 12]. So far, a few hundreds of genes have been identified, whose mutations and/or dysregulation of expressions were linked to increase risk of IBD and CRC [13-16]. However, genetic factors together can only explain limited proportions of the disease incidences [17-19]. Conversely, other factors are believed to be major contributors, especially gut microbes [12, 20-22]. The latter, along with metabolites and antibiotics that they 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 [23].

Fecal microbial dysbiosis in IBD and CRC have been observed, and subsequently utilized to generate predictive models for patient stratification and/or risk evaluation [22, 24-26]. For example, IBD patients showed a reduction of taxa from *Firmicutes* phylum and enrichment of pathogenic species [25, 27, 28]. Several studies showed the IBD subtypes, namely CD and UC had distinctive gut microbiota and metabolic profile, 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 gut contribute to the dysbiosis through affecting metabolic functions of bacteria [24, 30, 31].

However, binary models (i.e. models capable of distinguishing patients of a particular disease from controls) created for single disease may lead to misdiagnosis, on account of

some species 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 [25, 27]. 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 that whether we could distinguish one gut illness from others using solely the 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 and help us avoid biases from individual study [33]; moreover, the latest surveys about CRC via meta-analysis suggest the necessity in collecting metagenome data as much as possible to identify consistently altered microbes in certain state [34, 35].

In this research, we collected thirteen metagenomics datasets for common intestinal diseases known to have strong links to gut microbiota, including three, three, and seven datasets for CD, UC and CRC respectively, and performed meta-analysis to (1) determine disease-specific and consistent microbial alterations, (2) elucidate possible mechanisms underlying the altered species associated with different disease states, and (3) generate high-performance multi-class models using taxonomic and metabolic profiles for easier and better clinical applications.

Results

Meta-analyses of common intestinal diseases identified disease-specific and shared gut microbe and pathway markers

To determine which microbial features consistently alter in given enteropathy as comparing with controls, we collected and processed raw data from thirteen shot-gun sequencing datasets, consisting of three ulcerative colitis (UC), three Crohn's disease (CD) and seven colorectal cancer (CRC) datasets. We first assessed whether microbial diversity differed between cases and controls in each of the datasets. In all CRC datasets except PRJDB4176, significant differences of Shannon diversity between controls and CRCs were not found (Supplementary Figure 1a). By contrast, alpha diversity decreased generally in CD patients, while didn't display a stable trend in UC patients (Supplementary Figure 1b, c). We also found that the human DNA contents (HDCs), as calculated as the percentage of sequencing reads mapped to the human genome, were significantly higher in patients in all diseases (all blocked p-value < 0.001 in CRC, UC and CD separately; Supplementary Figure 2), consistent with previous studies [36] that high level of deciduous epithelial and/or blood cells could be found in stools of patients with IBD or CRC resulting from gut injury and quickening cell cycles [21, 30, 37].

We then performed meta-analysis on each disease to identify microbes that showed significantly differential abundances in at least two datasets of the same disease and referred them as "marker species" accordingly. We used univariate Wilcoxon rank sum test (see Methods) on relative abundance of species to identify differential species in each dataset (Supplementary Figure 3). Consequently, we identified in total 13, 47 and 16

marker species in UC, CD and CRC respectively; among which, ten were found in at least two diseases and one was common to all diseases (Figure 1a). For CRC, marker species were largely disease-enriched, including *Fusobacterium nucleatum*, *Parvimonas micra* and *Peptostreptococcus stomatis*, most of which were consistent with previous studies [34, 35]. Conversely, CD patients showed a depletion of control-enriched species, including *Akkermansia muciniphila* and a few within the *Bacteroides* genus, which were also in accordance with previous studies [25, 27, 28]. However, marker species in UC were a mix of both but mainly driven by disease-enriched microbes (Figure 1a), unlike a recent study that primarily showed a decreasing of control-enriched species [27]; the discrepancies are likely due to study-specific results (Figure 1a).

To further characterize these marker microbes, we first focused on patient-enriched ones and clustered them according to their distributions (i.e. prevalence) in all the case samples. To overcome the variances in species abundances in different studies, we defined positive samples for a given species as those in which the species was found with higher relative abundance than 95% controls. We obtained a binarized matrix with each row representing adisease-enriched marker microbe, and each column representing a patient; the contents of the matrix could be either “1”, indicating the presence of a species in the corresponding patient, or “0” indicating its absence. Calculating the Jaccard distance among the disease-enriched markers in the matrix and clustering species with ward.D2 algorithm, we obtained in total five clusters and classified them as disease-specific and mixed according to their marker species contents (Figure 1b). As we have expected, microbes in disease-specific clusters, especially that of UC and CRC,

showed highest prevalence and were significantly enriched in their respective diseases (Figure 1c). The results suggest that with the disease-specific clusters, it would be possible to stratify different diseases using microbial profiles; however, it is noteworthy that the prevalence was generally higher for marker microbes in the mixed clusters (Figure 1c, Mixed II & I) than those in the disease-specific clusters, highlighting the challenges in distinguishing different diseases.

We further revealed interesting patterns among the disease-enriched marker species based on their phylogenetic relationships (Figure 2). Firstly, disease-specific markers often form their own branches, and are mostly closely related on the phylogenic tree, i.e. they have rather low-levels of last common ancestors (LCAs) at the order, family and even genus levels (Figure 2). Secondly, the taxonomic composition of shared marker microbes is much more diverse; although they could form a single cluster on the phylogenetic tree (with the exception of *F. nucleatum*) and most of them belong to the *Firmicutes* phylum, they have an LCA of *Bacteria* at the Kingdom level. Lastly and most importantly, this phylogenetic tree revealed closer relationships between UC and CRC than to CD. For example, UC- and CRC-specific markers are next to each other on the tree and all belong to the order of *Bacteroidales*. Furthermore, they form a branch together with the shared markers (Mixed I & II), which is parallel to the branch formed by CD-specific markers (Figure 2). These results are consistent with the facts that UC patents are at higher risk for developing CRC, as compared with CD patients [9]. *F. nucleatum* was an outlier of the shared clusters (Mixed I & II): it was known as a CRC biomarker [34, 35], belonging to the Mixed I cluster, but was phylogenetically closer to the CD-specific markers (Figure 2); this

was likely due to the fact that *F. nucleatum* could be found in fecal samples with high prevalence in many diseases, including UC and CRC (Supplementary Figure 4, using data from GMrepo, a curated online database of human gut metagenomes [38]). Together, we revealed phylogenetic patterns of the marker microbes that could only be revealed through cross-disease analysis.

In addition to taxonomic profiles, we also processed the metabolic profiles predicted from metagenomics data using HuManN2 (see Methods), and identified in total 74 marker pathways with significantly different abundances between cases and controls in at least two datasets of each disease; among which eight were altered in two diseases (Supplementary Figure 5). Different from species level, there is no altered pathway common to all three enteropathies, suggesting the metabolic networks are more diverse in common gut illnesses. Consistent to previous results, most of the CD marker pathways were control-enriched; for example, pathways for amino acid biosynthesis, such as superpathway of L-lysine, L-threonine and L-methionine biosynthesisII, were depleted in CD patients. tRNA charging as common features in IBD is a process of producing charged tRNA which can carry an amino acid into a growing polypeptide; its depletion was consistent to a previous report that gut microbes in IBD patients reduced protein biosynthesis for nutrient absorption and toxicity production [11]. Conversely, the gut community in CRC showed distinct characteristics with the increased capacities on fatty acid and lipid biosynthesis, essential pathways for membrane biosynthesis of cancer cells [39].

Clustered marker microbes signified distinct metabolic capacities, especially in degradation

We next characterized the marker metabolic pathways. Interestingly, we found that most of the disease-altered pathways showed statistically correlations with the marker species. As shown in Figure 3 and Supplementary Figure 6, we were able to recaptulate the species clusters (including the control-enriched cluster identified in the previous sections) using their correlated metabolic capacities, especially in degradation (Figure 3). This was expected since most of the within-cluster microbes were phylogenetically close (Figure 2) and expected to have similar metabolic capacities.

We focused mostly on the degradation pathways (Figure 3) because the latter could better reflect the overall physio-metabolic properties of the gut environment in diseased states. For example, *Eubacterium ventriosum* and a few control-enriched species were positively correlated with most CRC-depleted pathways, such as starch degradation V, sucrose degradation III and D-galactose degradation V (Figure 3). These species are capable of fermentating general carbohydrates and producing butyrate, which are of anti-inflammatory effects in the gut [40]. 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 a 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 [41, 42]. Similarly, IBD-enriched *Coprobacillus spp.* had an association with creatinine degradationl, a vital function in supplying energy and keeping balance between ADP and ATP, indicating

disruption of gut mucosal barrier dynamics [43]. Besides, the CD-enriched cluster, in association with HDC, also correlated significantly with several pathways for energy production, including superpathway of glucose and xylose degradation, glucose and glucose-1-phosphate degradation, superpathway of fucose and rhamnose degradation and superpathway of methylglyoxal degradation (Figure 3). Of note, most of the pathways that related with the CD-specific cluster didn't associate with any UC-altered bacteria, indicating that although microbial and pathway profiles overlapped significantly, they could capture different aspects of the gut metagenomics. 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, in particular IBD subtypes.

The gut metabolic properties are known to be influenced by food and microbial activities. Recenlty, researchers also revealed that cells/metabolites derived from the human host, likely due to compromised intestinal barrier (CIB), can also influence the growth of individual bacterium and the gut microbes as a whole [25, 29]. CIB could lead to increased human DNA contents in the gut metagenomics. As expected, we found that HDCs were significantly elevated in cases of all projects (Supplementary Figures 2). Suprisingly, we found HDCs were also significantly correlated with most marker species in CD and CRC (Supplementary Figure 7). Most interestingly, *E. ventriosum*, the only control-enriched bacterial marker shared by all three diseases, was negatively correlated with HDCs in UC and CD (Supplementary Figure 8); 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 [25]. However, only two control-enriched species in UC had associations with HDCs (Supplementary Figure 7), implying the pathologies of CD and UC may be dissimilar. Together, our results suggested that the increased human derived contents could be an important contributor to the changes in local gut enviroment and hence the microbial profiles.

Marker species showed increased connectivity in diseases, presumably due to more stressed conditions

Having shown that the alterations in intestinal ecosystems could contribute to gut microbiota dysbiosis, we further explored the inter-relationships among the marker species within each physical condition. We speculated that the gut ecosystem under diseased states may represent more stressed conditions in which the growth of all microbes, especially the marker species, would be under stronger constraints and selection; consequently, we would expect to find stronger inter-relationships among the marker species.

We thus constructed inter-species networks separately for cases and controls for each disease based on pairwise correlations of the species abundances. We used Spearman rank correlation to calculate the correlation coefficients (rho values) and selected those relationships with fdr < 0.05 to perform meta-analysis (see Methods). As we expected, we found species in the cases were indeed connected more often than they

were in the controls of respective disease datasets (Figure 4). For example, we found 18 pairs of correlated marker species (meta fdr<0.05, see Methods) in CRC, increased from 12 in the controls (Figure 4a,d). Similarly, we found much more correlations among markers in cases than in controls in CD; notably, we found increases in both positively and negatively correlated species pairs (Figure 4b,d). In UC, we found 9 pairs of correlated marker species (meta fdr<0.05) in cases, increased from 5 in the controls (Figure 4c,d).

We also identified the potential powerful species in the networks using eigenvector centrality scores (ECSs) (Supplementary Figure 9). ECS served as assessment of node influence in a weighted network, measures the importance of the given node considering not only its connections with others, but also the connections of its related nodes. We found that *Eubacterium hallii*, *Anaerostipes hadrus*, *Lachnospiraceae bacterium 5 1 63FAA* and *Collinsella aerofaciens* were the pivotal species in controls of CD datasets, while in CD patients, *Ruminococcus obeum*, *Coprococcus comes*, *E. hallii* and *Eubacterium ramulus* took leading roles (Supplementary Figure 9). As expected, CRC-distinctive microbes, such as *Parvimonas spp* and *P. stomatis*, were the center in networks. Different from CD, nodes with highest eigenvector in UC were enriched in cases (Supplementary Figure 9).

Considering that the bacteria owning more correlations have the possibility to be hubs in disease condition, we calculated betweenness centrality scores (BCSs) standing for transmission capacity (Supplementary Figure10). We found that *Klebsiella pneumoniae* and *Escherichia coli* as pathogenic species were hubs in CD cases, both of which as popular targets in treatment, could produce lipopolysaccharide (LPS) to protect

themselves against host immune system and infect mucosa so that thrive in gut [44, 45]. Nevertheless, health-related species were also the hub bacteria in patients with IBD. These results support the view that we should be cautious to use antibiotic in IBD therapy, which may disrupt fragile connections among species exhausting in cases, and cause some bacteria failure of recovery [46]. In reference to stable interior couplings of CRC-relevant cluster, as well as marked relations between abundance of species cluster and human DNA contents (HDCs) detected in feces, we speculate that the growth of this cluster can be largely attributed to lesions of gastrointestinal tract.

Multi-class models capable of distinguishing distinct diseases using microbial and metabolic markers

We built multi-class 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 10-time 10 folds cross-validation for avoiding biases due to only one-time split (see Methods). The model based on combined markers in taxonomic and metabolic level reached the highest accuracy with true positive (TP) rate of 0.77 (Figure 5a). The classification errors were mostly contributed by UC samples, which share major markers with CD patients (Figure 1a, Supplementary Figure 5), and perhaps contain patients with inactive inflammation [25]. Nonetheless, we achieved slightly better classification performance (i.e. TPs) on UC and CD than a recent study (0.58 and 0.89 TRPs for UC and CD samples in our study, as compared to 0.49 and 0.66 in ref [25]).

We also built two additional models, used one to distinguish cases from controls (referred to as “binary” model), and the other to assign cases to distinct diseased states (referred to as “three-class” model). The latter was particularly important and clinically relevant due to the clinical overlap in presentation of these diseases, as well as the risk of IBD patients to eventually develop CRC. We evaluated the performance for the binary models on the 685 controls and 963 cases through 10-time 10-fold cross-validation, and found that all models performed well, with the area under the receiver operating characteristics curves (AUROC) scores ranging from 0.82 to 0.89. Notably, taxonomic based models in generally performed better than the metabolic based models, as the model with the combined profile attained a highest accuracy (Figure 5b). Surprisingly but expected, models using only the marker species/pathways performed comparable to those that used all species/pathways, especially the combined-based model (Figure 5b), suggesting that the much-shortened list of markers are of practical and clinical value.

Surprisingly, all three-class models again performed quite well according to the overall TP values (from 0.8 to 0.9) and much better than the four-class model (Figure 5a & c, Supplementary Figure 11). Contrary to the binary models, metabolic-based models out-performed taxonomic-based models (Supplementary Figure 11); the combined model of course performed the best with an overall accuracy of 0.9 (i.e. 90% of patients could be accurately diagnosed) (Supplementary Figure 11c). Again, we found that models using only the markers performed similarly to those using all the data, with an overall accuracy of 0.89 on the combined species and pathway markers (Figure 5c).

We noticed significant differences among diseases in terms of TP values in the

“three-class” models. In particular, we achieved high accuracies for CRC and CD (99% and 92% TP rates respectively) (Figure 5c), as compared with the relatively low TP rate for UC (74%). The latter was likely due to the fact that UC shared most of its markers with CD and had only few unique markers; consequently, most of the misclassified UC cases were predicted as CD (Figures 1a & 5, Supplementary Figure 5). These results suggested that additional information other than fecal metagenomics, such as physiological, genetic and clinical information of the human hosts are required to further improve the prediction accuracies.

Conclusions

In this study, we collected fecal metagenomics datasets for three common intestinal diseases, namely colorectal cancer (CRC), Crohn’s disease (CD) and ulcerative colitis (UC), totaling 11 projects, 13 datasets, 963 patients and 685 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 65 marker species and 74 marker pathways that were significantly changed in the same direction (i.e. case-depleted or case-enriched) in at least two independent studies of the same disease. We clustered the marker species into disease-specific and mixed clusters according to whether or not the member species were unique to a certain disease, and revealed their distinct phylogenetic relationships: for example, member species of the disease-specific clusters are often closely related and belong to the same genus or family

and can form a single clade on the NCBI taxonomic tree; conversely, although species of the mixed clusters can also form a single clade, they are much more diverse and have the last common ancestor of the phylum or kingdom level. Strikingly, UC- and CRC- specific species are phylogenetically closer to each other than to those of CD (Figure 2), in part due to the fact that UC-patients are at higher risk of developing CRC [9].

We then characterized the marker pathways. We first revealed that almost all marker pathways correlated significantly with marker species (Figure 3 & Supplementary Figure 6); additionally, clustered marker species tended to correlate significantly with the same sets of pathways (Figures 2&3). 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 human DNA contents (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 compromised intestinal barrier (CIB) (Supplementary Figure 2). The elevated HDCs in all datasets 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 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 as compared with the respective controls. We speculate that the diseased gut may represent a more stress environment due to the physio-metabolic changes including inflammation and/or bleeding; the inhabitant microbes are thus under

stronger selection, and show either more collaboration (positive correlation) or competition (negative correlation). Consequently, we found increased numbers of both positive and negative correlations among the marker species in the diseased samples for each disease.

Utilizing the identified marker species and pathways, we obtained three high-performance models for disease identification and patent stratification. The first "four-class" model could separate samples into controls (presumably healthy) or individual diseases (Figure 5a), with an overall true-positive rate (TPR) of 0.77. UC has the lowest TPR (0.58, Figure 5a) in this model; however, most of the wrongly classified samples went to CD, consistent with previous efforts [25] and the fact that UC had very few unique markers and shared most of its marker with CD (Figure 1a and Supplementary Figure 5). Regardless, it represents one of the best models that could classify IBD subtypes with TPR as 0.58 and 0.89 separately in UC and CD samples, while in a previous study TPR 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 [25]. We also built two additional models, including a "binary" model to distinguish cases from controls, and a "three-class" model to assign cases to distinct diseased states (referred to as "three-class" model). In our opinion, both are relevant in clinical applications. For example, the binary model, with AUROC of 0.88, can inform the subjects for further clinical inspections such as colonoscopy, while the "three-class" model, with an overall TPR of 0.89, can evaluate the UC and CD patient for potential CRC risks.

Taken together, our results demonstrated the necessity and feasibility of

metagenome-based multi-disease 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 on 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.

Methods

Data collection

Raw sequencing reads of the seven colorectal cancer and four inflammatory bowel disease datasets, were retrieved from European Nucleotide Archive (ENA) under the following ENA identifiers in order: PRJEB6070 [22], PRJEB27928 [34], PRJEB12449 [47], PRJEB10878 [26], PRJEB7774 [48], PRJDB4176 [34], PRJNA447983 (cohort1) [34], SRP057027 [24], PRJEB1220 [49], PRJNA400072 [25] and PRJNA389280 [28]. We obtained metadata from the corresponding studies. In total 863 patients and 675 controls were included in the meta-analysis. For detailed analyses, we split two IBD projects into two CD datasets and two UC datasets. In total we obtained three UC datasets, three CD datasets and seven CRC datasets. Details of the datasets and associated meta-data are provided in Supplementary Table 1.

Metagenomics data analysis

Low quality reads and adapters were removed via Trimmomatic (version 0.35) using the

Truseq3 adapter files and a MINLEN cutoff of 50 [50]. To detect and remove human DNA contents (HDCs), we aligned the remaining reads to the human reference genome (hg19) using bowtie2 (version 2.3.4.3) [51] with default settings, and computed HDC as the percentage of mapped reads out of total remained reads in each sample. The subsequent reads were then quantified in taxonomic and functional levels via using MetaPhlAn2 mapping to mpa_v20_m200 database and HUMAnN2 mapping to ChocoPhlAn database and full UniRef90 database [52, 53].

In each dataset, species with max abundance <0.1% in all samples and species whose average abundance across all samples below 0.01% were removed. Pathways with maximum relative abundances below 1×10^{-6} in all samples and with zero value in over 15% samples were removed. The abundance data were then loaded into R and analyzed (https://www.r-project.org; version 3.5.1).

Statistical analyses

Shannon index as overall composition in per dataset was calculated using diversity function in vegan package. Differences were estimated using Wilcoxon Rank Sum test. We took case-control comparisons on the fecal HDCs in each disease type and examined significances through the wilcox_test function blocked with “dataset” in coin package. Univariate analyses on the relative abundances of species and metabolic pathways were performed to detect significant differential features between cases and controls with a non-parametric Wilcoxon Rank Sum test in each dataset. p-values were corrected using p.adjust function with method=”fdr. Features with adjusted p-value < 0.05 in the same direction in at least two studies of the same disease were considered as consistently

disease-altered features. Spearman correlation was used to identify HDC-related microbial features, which maintained a positive or negative relationship with HDC (P-value < 0.05) in at least two datasets of each disease. To estimate prevalence of case-enriched species, we used the 95% percentile abundance of per species in all health as corresponding threshold, and defined a sample was species-positive if reaching this threshold. Using the binary prevalence matrix from all patients, we calculated the Jaccard distance among case-enriched species with the vegdist function and clustered them with the ward.D2 algorithm to visualize their co-occurrences. We then performed Cochran-Mantel-Haenszel test with "dataset" as the block object using cmh_test function, to compare ubiquity of case-abundant clusters between patients and controls in each disease type.

Construction of phylogenetic tree

The phylogenetic tree of case-enriched species was generated based on their NCBI taxonomy using an online tool, phyloT (https://phylot.biobyte.de/), setting internal nodes collapsed and polytomy as no. The tree file then was visualized using Evolview ver3, a webserver for annotation and management phylogenetic trees [54]. The nodes were colored depending on their corresponding clusters identified in previous section. The last common ancestors were determined according to the NCBI taxonomy of the species in corresponding branch.

Random forest classifiers

To classify cases and controls in the whole samples, we built random forest binary models using the repeated ten-fold cross-validation with randomForest and createMultiFolds

function in the randomForest package and caret package respectively, that duplicates ten-fold cross validation ten times to prevent biases due to one-time split. We computed the area under receiver operating characteristic (AUROC) as evaluation of binary classifiers based on diverse types of features, including altered species, total species, altered pathways and total pathways. Then we applied the similar method to establishing four-class and multi-case classification, for judging disease condition. Outcome and accuracy from multi-category classifiers were showed in the form of confusion matrix via the confusionMatrix function.

Microbial ecosystem analyses

Species-species correlations were estimated with Spearman's rank correlation separately in case and controls of each dataset. After adjusting p-values and extracting correlations with fdr < 0.05, we performed meta-analysis of correlation coefficients in distinct disease physical condition using random effect models with metacor function from meta package, which summarizes overall correlation based on Fisher's z transformation [33, 55]. The summary values were turned back into correlations for presentation. Results with corrected random p-values <0.05 were used to construct association networks. We evaluated eigenvector centrality and betweenness centrality of per network using correlation strength as weight with igraph package in R, and visualized networks in Cytoscape (version 3.7.2) [56]. The size of node indicated prevalence of the bacteria in counterpart samples. Positive and negative rectified coefficients as strength of edges were painted gray and red separately. We constructed the steady network from health based on connections occurred in at least 50% control datasets. In the same way, the

relations between altered species and metabolic pathways in each dataset were computed using Spearman's rank correlation, and those with fdr <0.05 were chosen to meta-analysis. Relative abundance for correlation were added pseudo values 1e-06 and 1e-09 respectively in taxonomic and functional levels. We preferred the correlations present at over 20% dataset.

Reference

1. Zhu, J., et al., *Epidemiological Trends in Colorectal Cancer in China: An Ecological Study.* Dig Dis Sci, 2017. **62**(1): p. 235-243.
2. Ye, Y., et al., *The epidemiology and risk factors of inflammatory bowel disease.* Int J Clin Exp Med, 2015. **8**(12): p. 22529-42.
3. Vuik, F.E., et al., *Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years.* Gut, 2019.
4. Araghi, M., et al., *Changes in colorectal cancer incidence in seven high-income countries: a population-based study.* Lancet Gastroenterol Hepatol, 2019.
5. M'Koma, A.E., *Inflammatory bowel disease: an expanding global health problem.* Clin Med Insights Gastroenterol, 2013. **6**: p. 33-47.
6. Ananthakrishnan, A.N., *Epidemiology and risk factors for IBD.* Nat Rev Gastroenterol Hepatol, 2015. **12**(4): p. 205-17.
7. Ng, S.C., et al., *Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies.* Lancet, 2018. **390**(10114): p. 2769-2778.

8. Stapley, S.A., et al., *Clinical features of bowel disease in patients aged <50 years in primary care: a large case-control study.* Br J Gen Pract, 2017. **67**(658): p. e336-e344.

9. Farraye, F.A., et al., *AGA Technical Review on the Diagnosis and Management of Colorectal Neoplasia in Inflammatory Bowel Disease.* Gastroenterology, 2010. **138**(2): p. 746-774.e4.

10. Abdalla, M. and H. Herfarth, *Rethinking Colorectal Cancer Screening in IBD, Is It Time to Revisit the Guidelines?* Journal of Crohn's and Colitis, 2018. **12**(7): p. 757-759.

11. Ni, J., et al., *Gut microbiota and IBD: causation or correlation?* Nat Rev Gastroenterol Hepatol, 2017. **14**(10): p. 573-584.

12. O'Keefe, S.J.D., *Diet, microorganisms and their metabolites, and colon cancer.* Nature Reviews Gastroenterology &Amp; Hepatology, 2016. **13**: p. 691.

13. Khor, B., A. Gardet, and R.J. Xavier, *Genetics and pathogenesis of inflammatory bowel disease.* Nature, 2011. **474**(7351): p. 307-317.

14. Liu, J.Z., et al., *Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.* Nature Genetics, 2015. **47**(9): p. 979-986.

15. Ma, X., B. Zhang, and W. Zheng, *Genetic variants associated with colorectal cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence.* Gut, 2014. **63**(2): p. 326.

16. Hu, Z., et al., *Quantitative evidence for early metastatic seeding in colorectal*

cancer. Nature Genetics, 2019. **51**(7): p. 1113-1122.

17. Frank, C., et al., *Concordant and discordant familial cancer: Familial risks, proportions and population impact.* Int J Cancer, 2017. **140**(7): p. 1510-1516.

18. Foulkes, W.D., *Inherited susceptibility to common cancers.* N Engl J Med, 2008. **359**(20): p. 2143-53.

19. Knights, D., et al., *Complex host genetics influence the microbiome in inflammatory bowel disease.* Genome Med, 2014. **6**(12): p. 107.

20. Lane, E.R., T.L. Zisman, and D.L. Suskind, *The microbiota in inflammatory bowel disease: current and therapeutic insights.* J Inflamm Res, 2017. **10**: p. 63-73.

21. Dickinson, B.T., et al., *Molecular markers for colorectal cancer screening.* Gut, 2015. **64**(9): p. 1485-94.

22. Zeller, G., et al., *Potential of fecal microbiota for early-stage detection of colorectal cancer.* Mol Syst Biol, 2014. **10**: p. 766.

23. Rooks, M.G. and W.S. Garrett, *Gut microbiota, metabolites and host immunity.* Nat Rev Immunol, 2016. **16**(6): p. 341-52.

24. Lewis, J.D., et al., *Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease.* Cell host & microbe, 2015. **18**(4): p. 489-500.

25. Franzosa, E.A., et al., *Gut microbiome structure and metabolic activity in inflammatory bowel disease.* Nat Microbiol, 2019. **4**(2): p. 293-305.

26. Yu, J., et al., *Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer.* Gut, 2017. **66**(1): p. 70.

27. Vich Vila, A., et al., *Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome.* Sci Transl Med, 2018. **10**(472).

28. Schirmer, M., et al., *Dynamics of metatranscription in the inflammatory bowel disease gut microbiome.* Nat Microbiol, 2018. **3**(3): p. 337-346.

29. Schirmer, M., et al., *Microbial genes and pathways in inflammatory bowel disease.* Nature Reviews Microbiology, 2019. **17**(8): p. 497-511.

30. Kim, M., et al., *Bacterial interactions with the host epithelium.* Cell Host Microbe, 2010. **8**(1): p. 20-35.

31. Zeng, M.Y., N. Inohara, and G. Nuñez, *Mechanisms of inflammation-driven bacterial dysbiosis in the gut.* Mucosal Immunology, 2017. **10**(1): p. 18-26.

32. Duvallet, C., et al., *Meta-analysis of gut microbiome studies identifies disease-specific and shared responses.* Nature Communications, 2017. **8**(1): p. 1784.

33. Hunter, J. and F. Schmidt, *Dichotomization of Continuous Variables: The Implications for Meta-Analysis.* Journal of Applied Psychology, 1990. **75**: p. 334-349.

34. Thomas, A.M., et al., *Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation.* Nat Med, 2019. **25**(4): p. 667-678.

35. Wirbel, J., et al., *Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer.* Nature Medicine, 2019. **25**(4): p.

679-689.

36. Lewis, James D., et al., *Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn’s Disease.* Cell Host & Microbe, 2015. **18**(4): p. 489-500.

37. Vincent, C., et al., *Excretion of Host DNA in Feces Is Associated with Risk of Clostridium difficile Infection.* J Immunol Res, 2015. **2015**: p. 246203.

38. Wu, S., et al., *GMrepo: a database of curated and consistently annotated human gut metagenomes.* Nucleic Acids Res, 2019.

39. Ong, E., et al., *Metabolic profiling in colorectal cancer reveals signature metabolic shifts during tumorigenesis.* Molecular & cellular proteomics : MCP, 2010.

40. Louis, P., G.L. Hold, and H.J. Flint, *The gut microbiota, bacterial metabolites and colorectal cancer.* Nature Reviews Microbiology, 2014. **12**(10): p. 661-672.

41. Colvin, H., et al., *Oncometabolite D-2-Hydroxyglurate Directly Induces Epithelial-Mesenchymal Transition and is Associated with Distant Metastasis in Colorectal Cancer.* Scientific Reports, 2016. **6**: p. 36289.

42. Han, J., et al., *Elevated d-2-hydroxyglutarate during colitis drives progression to colorectal cancer.* Proc Natl Acad Sci U S A, 2018. **115**(5): p. 1057-1062.

43. Kitzenberg, D., S.P. Colgan, and L.E. Glover, *Creatine kinase in ischemic and inflammatory disorders.* Clinical and Translational Medicine, 2016. **5**(1): p. 31.

44. Khan, S., et al., *Bacterial imbalance and gut pathologies: Association and contribution of E. coli in inflammatory bowel disease.* Crit Rev Clin Lab Sci, 2019. **56**(1): p. 1-17.

45. Kaur, C.P., J. Vadivelu, and S. Chandramathi, *Impact of Klebsiella pneumoniae in lower gastrointestinal tract diseases.* J Dig Dis, 2018. **19**(5): p. 262-271.

46. Ananthakrishnan, A.N., et al., *Environmental triggers in IBD: a review of progress and evidence.* Nat Rev Gastroenterol Hepatol, 2018. **15**(1): p. 39-49.

47. Vogtmann, E., et al., *Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing.* PloS one, 2016. **11**(5): p. e0155362-e0155362.

48. Feng, Q., et al., *Gut microbiome development along the colorectal adenoma–carcinoma sequence.* Nature Communications, 2015. **6**(1): p. 6528.

49. Le Chatelier, E., et al., *Richness of human gut microbiome correlates with metabolic markers.* Nature, 2013. **500**(7464): p. 541-6.

50. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

51. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

52. Segata, N., et al., *Metagenomic microbial community profiling using unique clade-specific marker genes.* Nature Methods, 2012. **9**(8): p. 811-+.

53. Franzosa, E.A., et al., *Species-level functional profiling of metagenomes and metatranscriptomes.* Nature Methods, 2018. **15**(11): p. 962-+.

54. Subramanian, B., et al., *Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees.* Nucleic Acids Res, 2019. **47**(W1): p. W270-W275.

55. Borenstein, M., et al., *A basic introduction to fixed-effect and random-effects models for meta-analysis.* Research Synthesis Methods, 2010. **1**(2): p. 97-111.

56. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

Funding

This work was partly supported by National Key Research and Development Program of China 2018YFC0910502 (to W.H.C.), National Natural Science Foundation of China (61932008, 61772368, 61572363), National Key R&D Program of China (2018YFC0910500), Natural Science Foundation of Shanghai (17ZR1445600), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01) and ZJLab.

Author contributions

W.H.C. and X.M.Z. designed the study. P.J. collected and analyzed the data, S.W. coordinated the data downloads and analysis. W.H.C., and P.J. wrote the manuscript with all authors contributing to the writing and providing feedbacks. All authors read and approved the final version of the manuscript.

Acknowledgements

Special thanks are due to Lei Liu, Xinming Li, and Na L Gao for valuable discussion.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.H.C. and X.M.Z..

Figures legends

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) in at least two datasets of corresponding disease are shown (fdr < 0.05, Wilcoxon rank sum test, Benjamini-Hochberg FDR correction), with their phylum shown on top. Red indicates case-enriched species, and blue indicates control-enriched one. **b.** Clustering of case-enriched microbes according to their Jaccard similarity of prevalence in all cases. The clustering was generated using the hclust function of R with the ward.D2 algorithm (see Methods); red indicates higher similarity of prevalence and white indicates higher divergence. The five clusters were named according to their members; UC-, CD- and CRC- clusters only contained disease-specific markers, while Mixed I and II contained markers from at least two diseases. **c**. The barplot shows the fraction of cases that are "positive" for given clusters in 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 Crochran-Mantel-Haenszel test with 'dataset' as the stratified factor; the 'star' symbols on the bars were used to indicate that marker species of a given cluster were significantly more prevalent in cases than in their corresponding controls (*: $p < 0.05$, **: $p < 0.01$).

Fig. 2 | Phylogenetic relationships of disease-specific and shared marker species.

Phylogenetic relationships of the marker species were based on the NCBI common tree, generated using the online phyloT tool (https://phylot.biobyte.de), and visualized using Evolview ver3.0 [54] (with manual annotations). The last common ancestors were determined according to the NCBI taxonomy information of the species in corresponding branch.

Fig. 3 | Marker microbes signified distinct degradation preferences. Shown here are the correlations in relative abundances between the degredation pathways and the marker species. Correlations with adjusted P-value below 0.05 in at least three datasets were selected to estimate in meta-analysis with random effects model (Spearman's rank correlation, Benjamini-Hochberg FDR correction). The pathways were clustered using 'complete' 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 block indicates positive correlation and blue indicates negative one. The asterisk indicates *P*-value of the corrected coefficient is below 0.01. The similar plot for biosynthesis- pathways is shown in Supplementary Figures 6.

Fig. 4 | Increased correlations of marker species in diseased conditions. **a-c**, Correlation networks among differential species of per state in corresponding disease datasets (Spearman's rank correlation and meta-analysis, see Methods). Color of nodes means the group that the species was clustered in as has been noted, and the sizes mean the prevalence of the bacteria in the overall health (or cases) of given datasets. Gray edges

indicate positive relationship and red edges indicate negative. Thickness of edges indicates correlation strength. **d,** statistics on the increases of positive correlated pairs in cases (red) as compared with controls (blue) in each of the three diseases. Y-axis shows the case-to-control ratios of each disease, with the controls being normalized to 1. The numbers in the bars indicate the actual numbers of correlated pairs. Please note only the positively correlated pairs are shown here, because only CD patients contain negatively correlated marker species (see the main text).

Fig. 5 | Random forest models for patient stratification using taxonomic and/or metabolic profiles. **a.** Confusion matrix evaluation of the four-class model based on combined marker features for distinguishing different physical condition. The number in row i and column j in the matrix on the left part indicates how many patients with disease i actually were categorized to disease j in model. Color filling the cell means the relative size of the number in the corresponding row. The right part is the true positive rate (TPR) for per disease type. Total accuracy indicates the fraction of all correct predictions, and 95% CI is the confidence interval of accuracy. **b.** Binary models for distinguishing cases from controls were constructed using various type of features as shown in plot. Each bar represents accuracy of corresponding model. AUROC as accuracy in axis x is evaluated through ten-time ten-fold random forest cross-validation in all samples. The solid and dashed vertical lines indicate AUROC values of 0.90 and 0.85 respectively. **c.** Assessment of three-class model for partioning disease conditions. Only cases were used. Shown here is model built with combined marker features; models using all species and/or

pathways and individual marker species or pathways are shown in Supplementary Figure 11 separately. Performance of the model was displayed as a confusion matrix like Figure 5a.

Additional files

Supplementary Table 1. A list of CRC and IBD datasets used in the study and numbers of controls and cases.

Supplementary Figure 1 | Shannon diversity between cases and controls in each dataset. Alpha diversity measured as Shannon diversity was calculated for overall gut microbes (n = 363). P-value was computed in each dataset using Wilcoxon rank sum test, and showed as significant label (ns : P-value ≥ 0.05; * : P-value < 0.05; ** : P-value < 0.01). **a**, CRC; **b**, CD; **c**, UC.

Supplementary Figure 2 | All datasets were found to contain elevated human DNA contents (HDCs) in cases as compared with their corresponding controls. The P-value of each dataset was computed using a univariate Wilcoxon rank sum test, and the overall P-value of each disease type was calculated using a blocked Wilcoxon test with 'dataset' as the block object.

Supplementary Figure 3 | Generalized fold changes of identified marker microbes that

showed consistent changes in at least two datasets of the respective diseases. Generalized fold changes were calculated through taking the mean of the fold change between cases and controls at per quantile (from 0.1 to 0.9, 0.2 as step length). The identified markers were colored according to the datasets in which the species altered evidently, and were shaped depending on the corresponding disease type of datasets.

Supplementary Figure 4 | Prevalence of *F. nucleatum* in fecal samples of selected diseases. Data were obtained from GMrepo, a curated online database of human gut metagenomes [38]. GMrepo contained relative abundance profiles of gut fecal microbes in totally 93 diseases. We selected diseases that have at least 100 samples in the database, and calculated the prevalence of *F. nucleatum* for each disease; here the prevalence of *F. nucleatum* in a disease is defined as the percentage of samples that contains this bacterium with relative abundance higher than 0.01%.

Supplementary Figure 5 | Identification of marker pathways that showed consistent changes in at least two datasets of the respective diseases. The relative abundances of pathways were identified using HuManN2 (see Methods) on the metagenomics data. Pathways significantly enriched in cases (or controls) in at least two datasets of corresponding disease are shown (fdr < 0.05, Wilcoxon rank sum test, Benjamini-Hochberg FDR correction). Red indicates case-enriched pathway, and blue indicates control-enriched one.

Supplementary Figure 6 | Marker microbes showed distinct biosynthesis preferences. Similar to Figure 3, however, shown here are the correlations in relative abundances between the biosynthesis pathways and the marker species. Correlations with adjusted P-value below 0.05 in at least three datasets were selected to estimate in meta-analysis with random effects model (Spearman's rank correlation, Benjamini-Hochberg FDR correction). The pathways were clustered using 'complete' 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 block indicates positive correlations and blue indicates negative ones. The asterisk indicates *P*-value of the corrected coefficient is below 0.01.

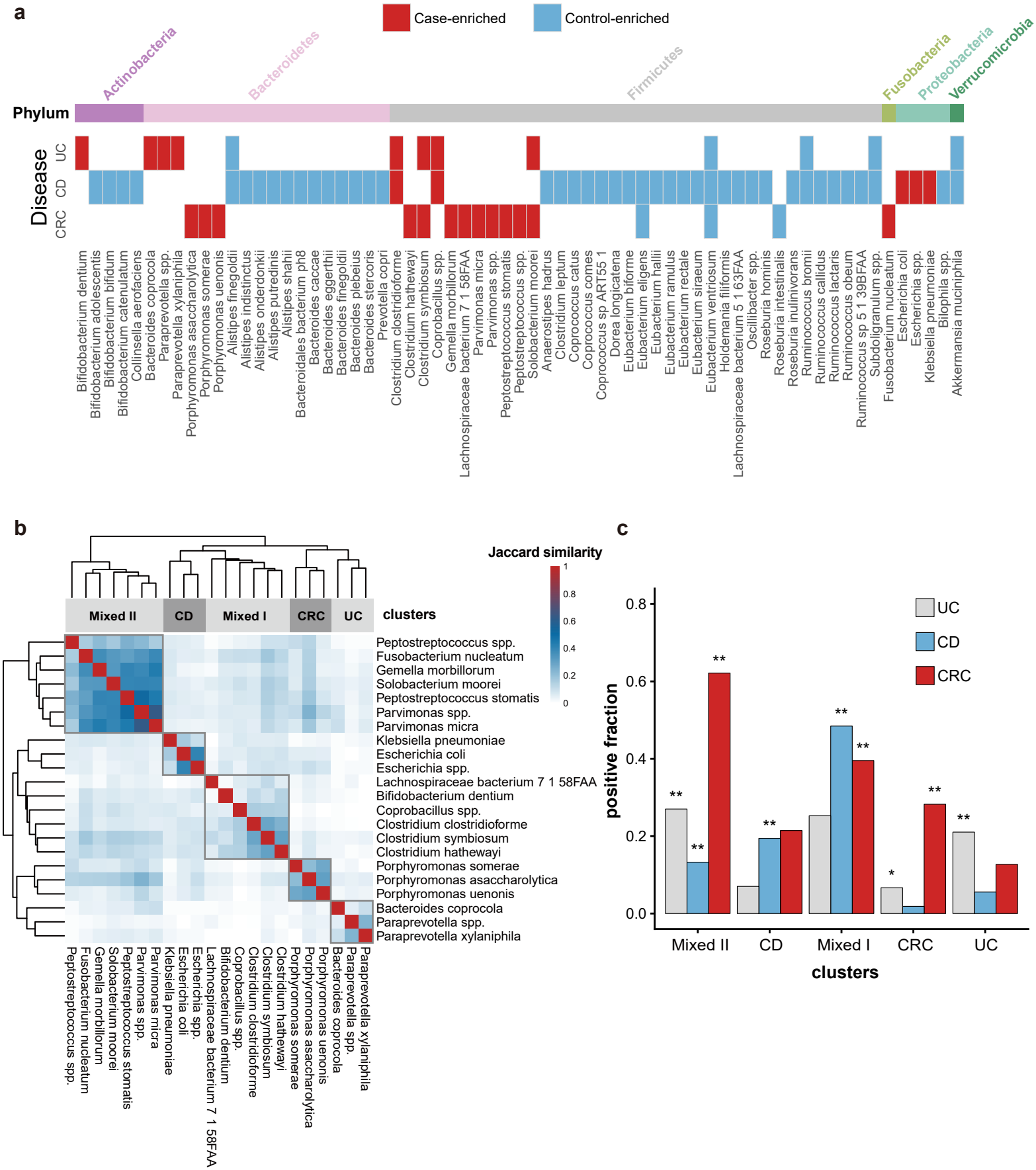
Supplementary Figure 7 | Most marker species correlated significantly with human DNA contents. Each block indicates the marker species found in at least two datasets of corresponding disease. Red means that the species was HDC-related evidently in at least two datasets of corresponding disease type. Blue means the species was not HDC-related.

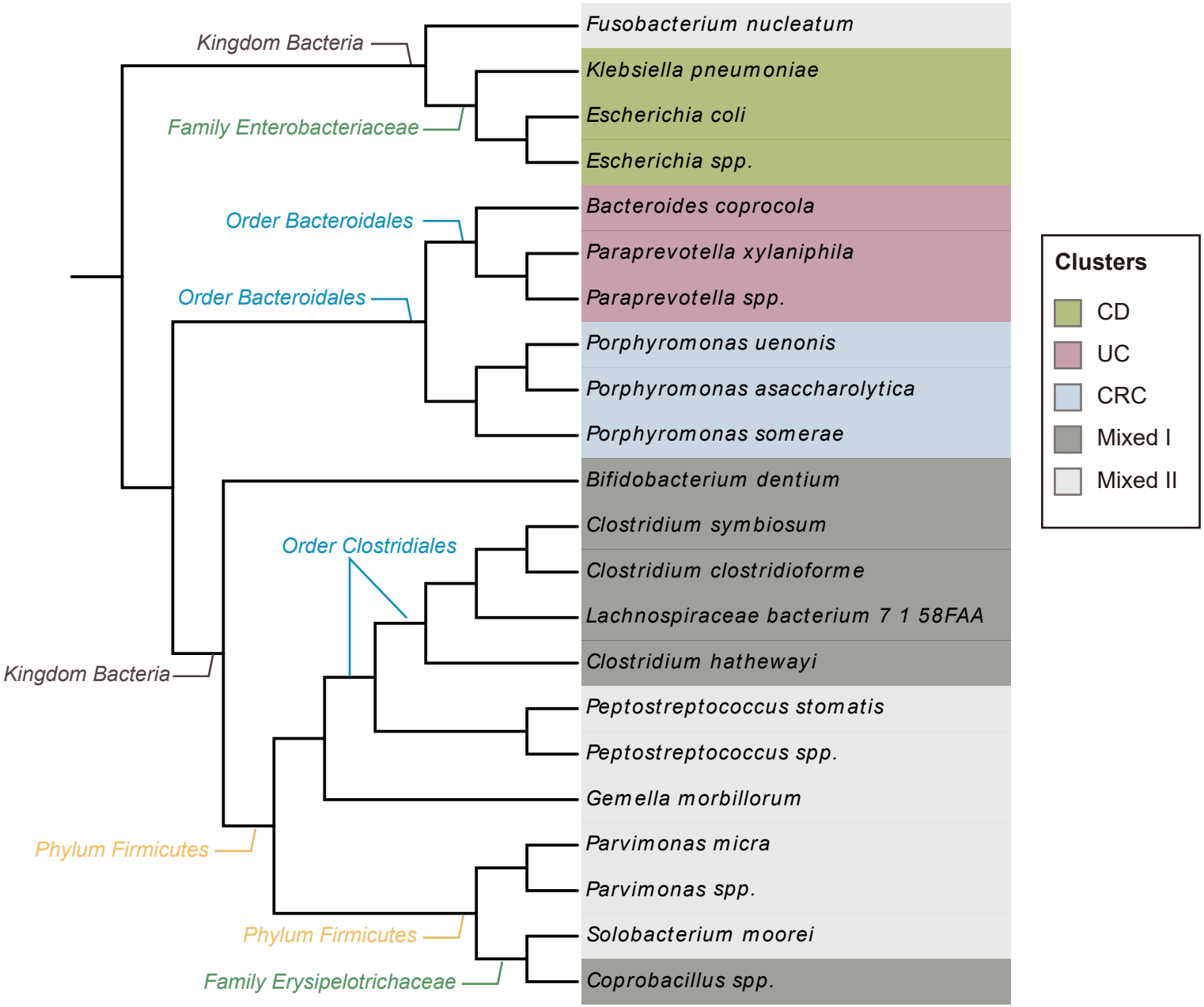
Supplementary Figure 8 | *Eubacterium ventriosum* was negatively correlated with HDCs in UC and CD patients, but not CRC. Correlation between relative abundance of *E. ventriosum* and HDC in each disease type was computed using Spearman correlation. Gray points indicate the controls in corresponding disease type.

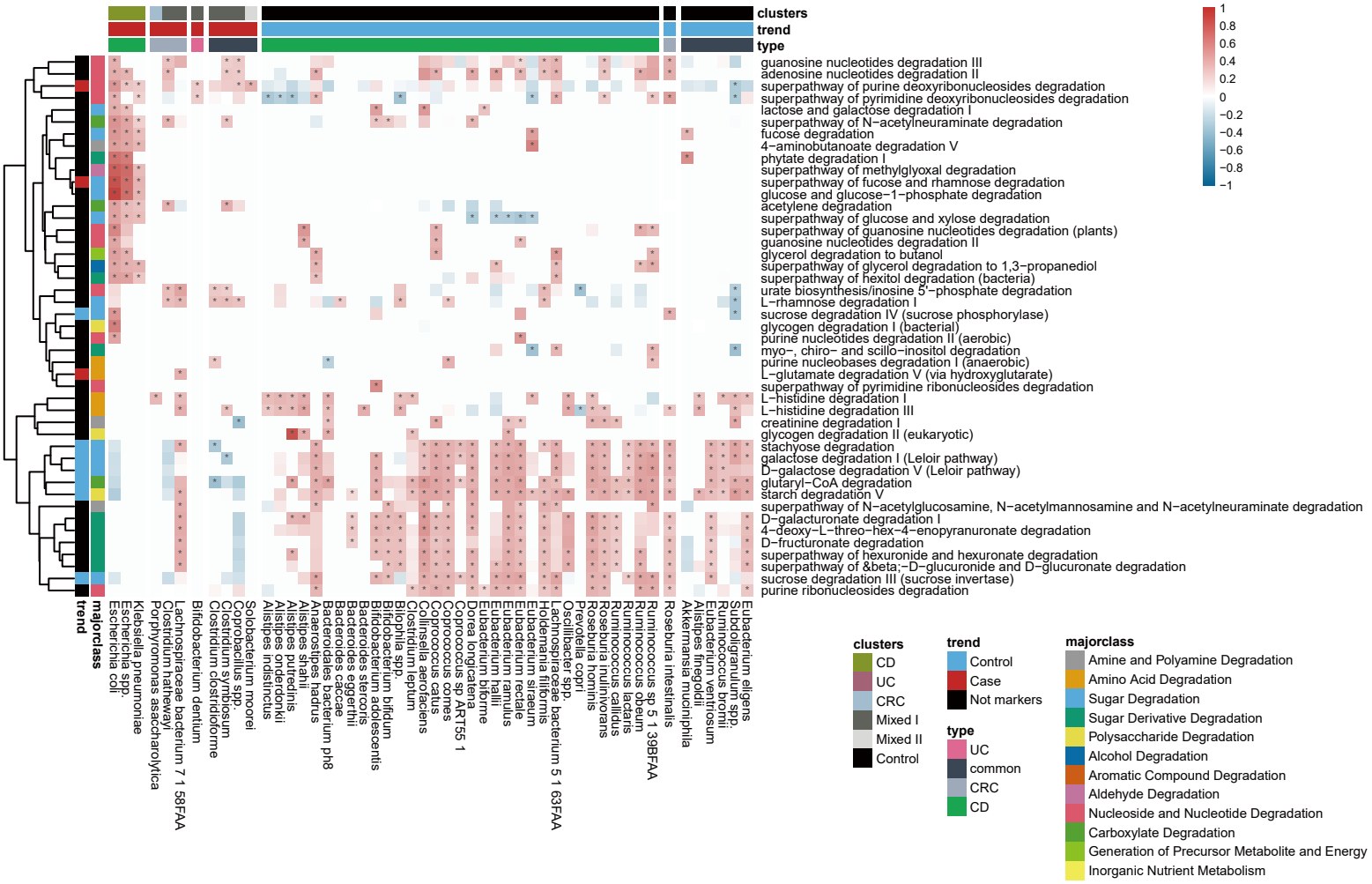
Supplementary Figure 9 | Evaluation for eigenvector centralities of networks showed in Figure 4. Blue bars indicate the **eigenvector** centralies of networks in health from datasets for a given disease, and red bars indicate **eigenvector** centrailies in networks of patients.

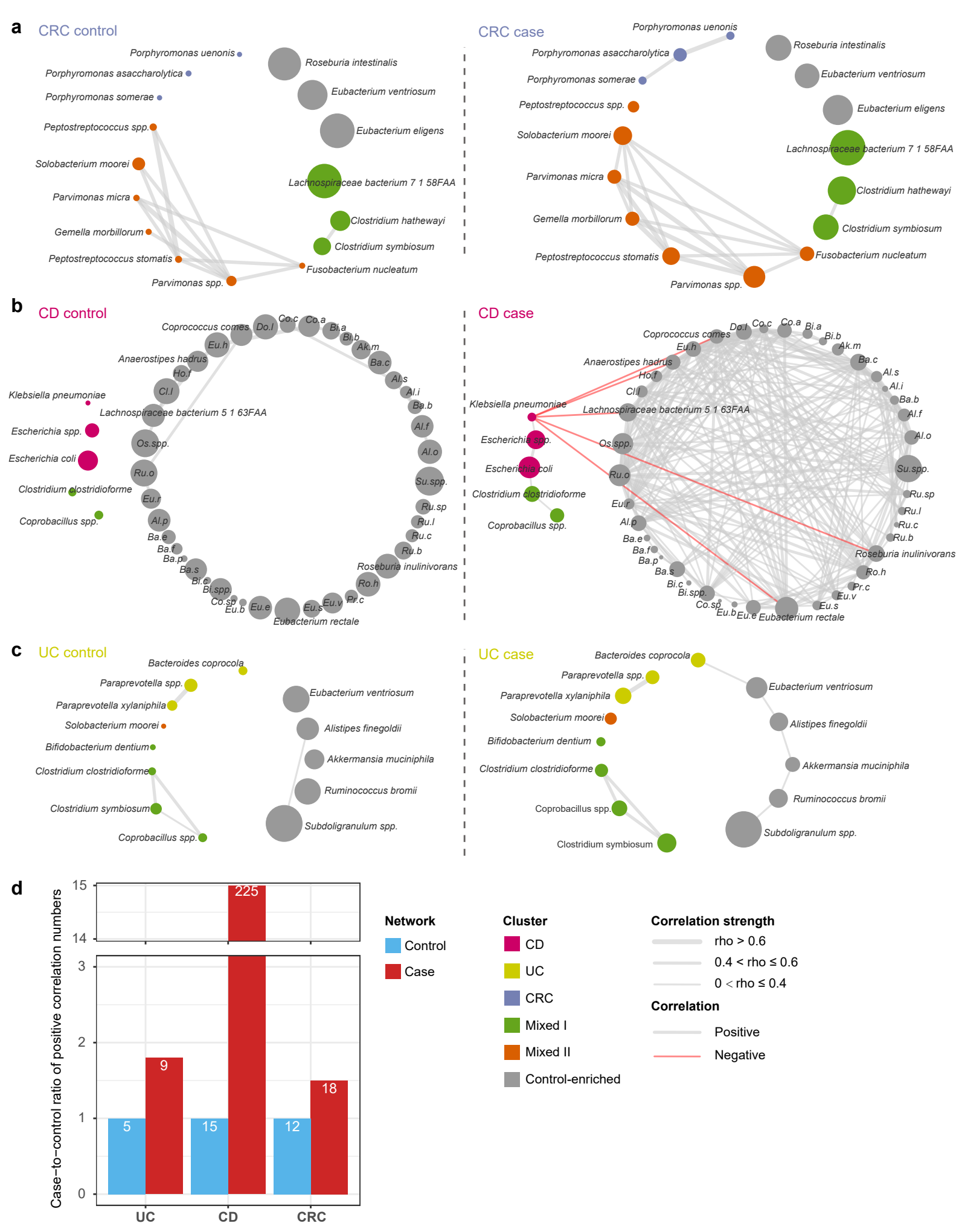
Supplementary Figure 10 | Evaluation for betweenness centralities of networks showed in Figure 4. Blue bars indicate the betweenness centralies of networks in health from datasets for a given disease, and red bars indicate betweenness centrailies in networks of patients.

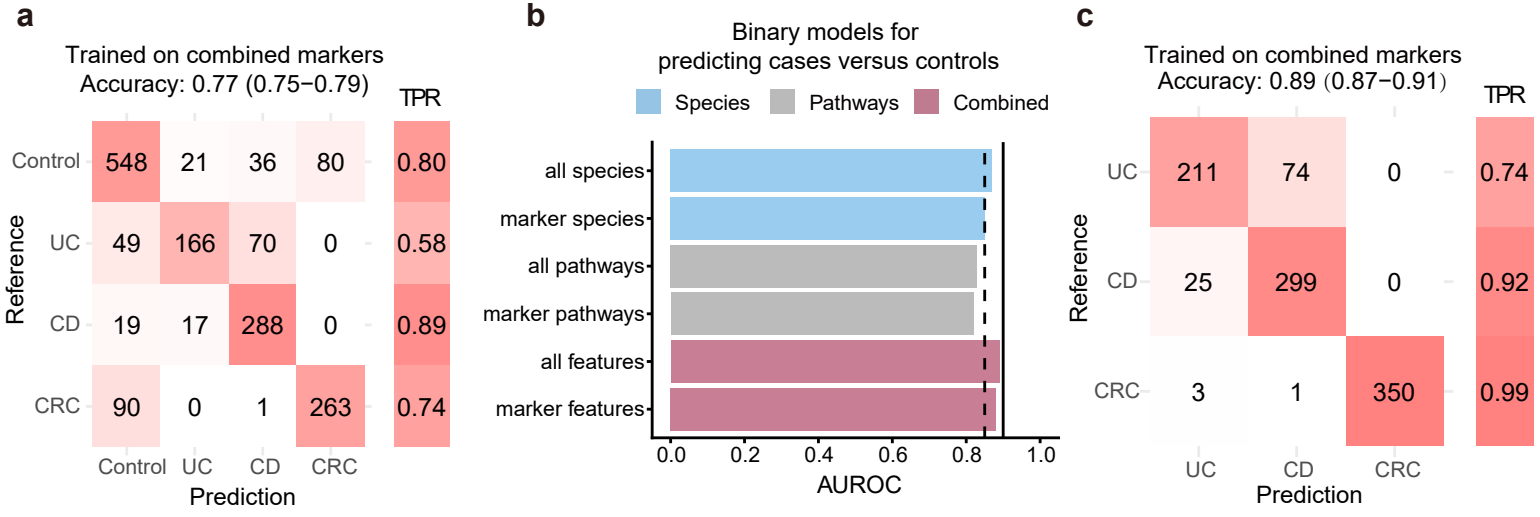
Supplementary Figure 11 | Multi-disease models based on various features. **a-c**: taxonomic-, metabolic-, and combined profile; **d-e**: marker species and marker pathways. Confusion matrix evaluations of multi-cases models based on various types of features for distinguishing different physical condition. The number in row i and column j in the matrix on the left part indicates how many patients with disease i actually were categorized to disease j in model. Color filling the cell means the relative size of the number in the corresponding row. The right part is the true positive rate for per disease type. Total accuracy indicates the fraction of all correct predictions, and 95% CI is the confidence interval of accuracy.


a
Case-enriched
Control-enriched
Phylum
Actinobacteria
Bacteroidetes
Firmicutes
Fusobacteria
Proteobacteria
Verrucomicrobia
Disease
UC
CD
CRC
Bifidobacterium dentium
Bifidobacterium adolescentis
Bifidobacterium bifidum
Bifidobacterium catenulatum
Collinsella aerofaciens
Bacteroides coprocola
Paraprevotella spp.
Paraprevotella xylaniphila
Porphyromonas asaccharolytica
Porphyromonas somerae
Porphyromonas uenonis
Alistipes finegoldii
Alistipes indistinctus
Alistipes onderdonkii
Alistipes putredinis
Alistipes shahii
Bacteroidales bacterium ph8
Bacteroides caccae
Bacteroides eggerthii
Bacteroides finegoldii
Bacteroides plebeius
Bacteroides stercoris
Prevotella copri
Clostridium clostridioforme
Clostridium hathewayi
Clostridium symbiosum
Coprobacillus spp.
Gemella morbillorum
Lachnospiraceae bacterium 7 1 58FAA
Parvimonas micra
Parvimonas spp.
Peptostreptococcus stomatis
Peptostreptococcus spp.
Solobacterium moorei
Anaerostipes hadrus
Clostridium leptum
Coprococcus catus
Coprococcus comes
Coprococcus sp ART55 1
Dorea longicatena
Eubacterium biforme
Eubacterium eligens
Eubacterium hallii
Eubacterium ramulus
Eubacterium rectale
Eubacterium siraeum
Eubacterium ventriosum
Holdemania filiformis
Lachnospiraceae bacterium 5 1 63FAA
Oscillibacter spp.
Roseburia hominis
Roseburia intestinalis
Roseburia inulinivorans
Ruminococcus bromii
Ruminococcus callidus
Ruminococcus lactaris
Ruminococcus obeum
Ruminococcus sp 5 1 39BFAA
Subdoligranulum spp.
Fusobacterium nucleatum
Escherichia coli
Escherichia spp.
Klebsiella pneumoniae
Bilophila spp.
Akkermansia muciniphila
b
Jaccard similarity
1
0.8
0.6
0.4
0.2
0
Mixed II
CD
Mixed I
CRC
UC
clusters
Peptostreptococcus spp.
Fusobacterium nucleatum
Gemella morbillorum
Solobacterium moorei
Peptostreptococcus stomatis
Parvimonas spp.
Parvimonas micra
Klebsiella pneumoniae
Escherichia coli
Escherichia spp.
Lachnospiraceae bacterium 7 1 58FAA
Bifidobacterium dentium
Coprobacillus spp.
Clostridium clostridioforme
Clostridium symbiosum
Clostridium hathewayi
Porphyromonas somerae
Porphyromonas asaccharolytica
Porphyromonas uenonis
Bacteroides coprocola
Paraprevotella spp.
Paraprevotella xylaniphila
c
UC
CD
CRC
positive fraction
0.8
0.6
0.4
0.2
0.0
**
*
Mixed II
CD
Mixed I
CRC
UC
clusters


Kingdom Bacteria
Family Enterobacteriaceae
Order Bacteroidales
Order Bacteroidales
Order Clostridiales
Kingdom Bacteria
Phylum Firmicutes
Phylum Firmicutes
Family Erysipelotrichaceae
Fusobacterium nucleatum
Klebsiella pneumoniae
Escherichia coli
Escherichia spp.
Bacteroides coprocola
Paraprevotella xylaniphila
Paraprevotella spp.
Porphyromonas uenonis
Porphyromonas asaccharolytica
Porphyromonas somerae
Bifidobacterium dentium
Clostridium symbiosum
Clostridium clostridioforme
Lachnospiraceae bacterium 7 1 58FAA
Clostridium hathewayi
Peptostreptococcus stomatis
Peptostreptococcus spp.
Gemella morbillorum
Parvimonas micra
Parvimonas spp.
Solobacterium moorei
Coprobacillus spp.
Clusters
CD
UC
CRC
Mixed I
Mixed II


clusters
trend
type
guanosine nucleotides degradation III
adenosine nucleotides degradation II
superpathway of purine deoxyribonucleosides degradation
superpathway of pyrimidine deoxyribonucleosides degradation
lactose and galactose degradation I
superpathway of N−acetylneuraminate degradation
fucose degradation
4−aminobutanoate degradation V
phytate degradation I
superpathway of methylglyoxal degradation
superpathway of fucose and rhamnose degradation
glucose and glucose−1−phosphate degradation
acetylene degradation
superpathway of glucose and xylose degradation
superpathway of guanosine nucleotides degradation (plants)
guanosine nucleotides degradation II
glycerol degradation to butanol
superpathway of glycerol degradation to 1,3−propanediol
superpathway of hexitol degradation (bacteria)
urate biosynthesis/inosine 5'−phosphate degradation
L−rhamnose degradation I
sucrose degradation IV (sucrose phosphorylase)
glycogen degradation I (bacterial)
purine nucleotides degradation II (aerobic)
myo−, chiro− and scillo−inositol degradation
purine nucleobases degradation I (anaerobic)
L−glutamate degradation V (via hydroxyglutarate)
superpathway of pyrimidine ribonucleosides degradation
L−histidine degradation I
L−histidine degradation III
creatinine degradation I
glycogen degradation II (eukaryotic)
stachyose degradation
galactose degradation I (Leloir pathway)
D−galactose degradation V (Leloir pathway)
glutaryl−CoA degradation
starch degradation V
superpathway of N−acetylglucosamine, N−acetylmannosamine and N−acetylneuraminate degradation
D−galacturonate degradation I
4−deoxy−L−threo−hex−4−enopyranuronate degradation
D−fructuronate degradation
superpathway of hexuronide and hexuronate degradation
superpathway of β−D−glucuronide and D−glucuronate degradation
sucrose degradation III (sucrose invertase)
purine ribonucleosides degradation
trend
majorclass
Escherichia coli
Escherichia spp.
Klebsiella pneumoniae
Porphyromonas asaccharolytica
Lachnospiraceae bacterium 7 1 58FAA
Clostridium hathewayi
Bifidobacterium dentium
Clostridium clostridioforme
Clostridium symbiosum
Coprobacillus spp.
Solobacterium moorei
Alistipes indistinctus
Alistipes onderdonkii
Alistipes putredinis
Alistipes shahii
Anaerostipes hadrus
Bacteroidales bacterium ph8
Bacteroides caccae
Bacteroides eggerthii
Bacteroides stercoris
Bifidobacterium adolescentis
Bifidobacterium bifidum
Blautia spp.
Clostridium leptum
Collinsella aerofaciens
Coprococcus catus
Coprococcus comes
Coprococcus sp ART55 1
Dorea longicatena
Eubacterium biforme
Eubacterium hallii
Eubacterium siraeum
Eubacterium rectale
Holdemania filiformis
Lachnospiraceae bacterium 5 1 63FAA
Oscillibacter spp.
Prevotella copri
Roseburia hominis
Roseburia inulinivorans
Ruminococcus callidus
Ruminococcus lactaris
Ruminococcus obeum
Ruminococcus sp 5 1 39BFAA
Roseburia intestinalis
Akkermansia muciniphila
Alistipes finegoldii
Eubacterium ventriosum
Ruminococcus bromii
Subdoligranulum spp.
Eubacterium eligens
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
clusters
CD
UC
CRC
Mixed I
Mixed II
Control
trend
Control
Case
Not markers
type
UC
common
CRC
CD
majorclass
Amine and Polyamine Degradation
Amino Acid Degradation
Sugar Degradation
Sugar Derivative Degradation
Polysaccharide Degradation
Alcohol Degradation
Aromatic Compound Degradation
Aldehyde Degradation
Nucleoside and Nucleotide Degradation
Carboxylate Degradation
Generation of Precursor Metabolite and Energy
Inorganic Nutrient Metabolism


a
CRC control
Porphyromonas uenonis
Porphyromonas asaccharolytica
Porphyromonas somerae
Peptostreptococcus spp.
Solobacterium moorei
Parvimonas micra
Gemella morbillorum
Peptostreptococcus stomatis
Parvimonas spp.
Roseburia intestinalis
Eubacterium ventriosum
Eubacterium eligens
Lachnospiraceae bacterium 7 1 58FAA
Clostridium hathewayi
Clostridium symbiosum
Fusobacterium nucleatum
CRC case
Porphyromonas uenonis
Porphyromonas asaccharolytica
Porphyromonas somerae
Peptostreptococcus spp.
Solobacterium moorei
Parvimonas micra
Gemella morbillorum
Peptostreptococcus stomatis
Parvimonas spp.
Roseburia intestinalis
Eubacterium ventriosum
Eubacterium eligens
Lachnospiraceae bacterium 7 1 58FAA
Clostridium hathewayi
Clostridium symbiosum
Fusobacterium nucleatum
b
CD control
Coprococcus comes
Anaerostipes hadrus
Klebsiella pneumoniae
Escherichia spp.
Escherichia coli
Clostridium clostridioforme
Coprobacillus spp.
Lachnospiraceae bacterium 5 1 63FAA
Roseburia inulinivorans
Eubacterium rectale
CD case
Coprococcus comes
Anaerostipes hadrus
Klebsiella pneumoniae
Escherichia spp.
Escherichia coli
Clostridium clostridioforme
Coprobacillus spp.
Lachnospiraceae bacterium 5 1 63FAA
Roseburia inulinivorans
Eubacterium rectale
c
UC control
Bacteroides coprocola
Paraprevotella spp.
Paraprevotella xylaniphila
Solobacterium moorei
Bifidobacterium dentium
Clostridium clostridioforme
Clostridium symbiosum
Coprobacillus spp.
Eubacterium ventriosum
Alistipes finegoldii
Akkermansia muciniphila
Ruminococcus bromii
Subdoligranulum spp.
UC case
Bacteroides coprocola
Paraprevotella spp.
Paraprevotella xylaniphila
Solobacterium moorei
Bifidobacterium dentium
Clostridium clostridioforme
Coprobacillus spp.
Clostridium symbiosum
Eubacterium ventriosum
Alistipes finegoldii
Akkermansia muciniphila
Ruminococcus bromii
Subdoligranulum spp.
d
Case−to−control ratio of positive correlation numbers
15
14
3
2
1
0
225
9
5
15
18
12
UC
CD
CRC
Network
Control
Case
Cluster
CD
UC
CRC
Mixed I
Mixed II
Control-enriched
Correlation strength
rho > 0.6
0.4 < rho ≤ 0.6
0 < rho ≤ 0.4
Correlation
Positive
Negative


a
Trained on combined markers
Accuracy: 0.77 (0.75−0.79)
TPR
Reference
Control
UC
CD
CRC
548 21 36 80 0.80
49 166 70 0 0.58
19 17 288 0 0.89
90 0 1 263 0.74
Control UC CD CRC
Prediction
b
Binary models for
predicting cases versus controls
Species
Pathways
Combined
all species
marker species
all pathways
marker pathways
all features
marker features
0.0 0.2 0.4 0.6 0.8 1.0
AUROC
c
Trained on combined markers
Accuracy: 0.89 (0.87−0.91)
TPR
Reference
UC
CD
CRC
211 74 0 0.74
25 299 0 0.92
3 1 350 0.99
UC CD CRC
Prediction