1 Metagenomic analysis of common intestinal diseases 2 reveals relationships among microbial signatures and 3 powers multi-disease diagnostic models

4

5 Puzi Jiang ^{1,6}, Sicheng Wu ^{1,6}, Qibin Luo ², Xing-ming Zhao ^{3,4,§}, Wei-Hua Chen

6 ^{1,5 §}

7

8 Affiliations:

9 ¹ Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei

10 Key Laboratory of Bioinformatics and Molecular-imaging, Center for Artificial

11 Intelligence Biology, Department of Bioinformatics and Systems Biology,

12 College of Life Science and Technology, Huazhong University of Science and

13 Technology, 430074 Wuhan, Hubei, China

14 ² Department of Genome Oriented Bioinformatics, Technische Universität

15 München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany

16 ³ Institute of Science and Technology for Brain-Inspired Intelligence, Fudan

17 University, Shanghai 200433, China

18 ⁴ Key Laboratory of Computational Neuroscience and Brain-Inspired

19 Intelligence, Ministry of Education, China

20 ⁵ College of Life Science, HeNan Normal University, 453007 Xinxiang, Henan,

21 China

1 ⁶ These authors contributed equally

2 [§] To whom correspondence should be addressed

3 Emails:

4 Puzi Jiang: puzjiang@hust.edu.cn

5 Sicheng Wu: wu_sicheng@hust.edu.cn

6 Qibin Luo: qibinluo@icloud.com

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

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

9

10 ABSTRACT

11 Common intestinal diseases such as Crohn's disease (CD), ulcerative colitis
 12 (UC) and colorectal cancer (CRC), share clinical symptoms and altered gut
 13 microbes, necessitating cross-disease comparisons and the use of
 14 multi-disease models. Here, we performed meta-analyses on thirteen fecal
 15 metagenome datasets of the three diseases. We identified 87 species and 65
 16 pathway markers that were consistently changed in multiple datasets of the
 17 same diseases. According to their overall trends, we grouped the
 18 disease-enriched marker species into disease-specific and -common clusters,
 19 and revealed their distinct phylogenetic relationships: species in CD-specific
 20 cluster are phylogenetically related, while those in CRC-specific cluster are
 21 more distant; strikingly, UC- specific species are phylogenetically closer to
 22 CRC, likely because UC-patients have higher risk of CRC. Consistent to their

1 phylogenetic relationships, marker species had similar within-cluster and
 2 different between-cluster metabolic preferences. There were part of marker
 3 species and pathways correlated with an indicator of leaky gut, suggesting a
 4 link between gut dysbiosis and human derived contents. Marker species
 5 showed more coordinated changes and tighter inner-connections in cases
 6 than the controls, suggesting that the diseased gut may represent a stressed
 7 environment and pose stronger selection to gut microbes. With the marker
 8 species and pathways, we constructed four high-performance (including
 9 multi-disease) models with AUROC of 0.87 and true positive rates up to 90%,
 10 and explained their putative clinical applications. We identified consistent
 11 microbial alterations in common intestinal diseases, revealed metabolic
 12 capacities and the relationships among marker bacteria in distinct states, and
 13 supported the feasibility of metagenome-derived multi-disease diagnosis.

14

15 Importance

16 Gut microbes have been identified as potential markers in distinguishing
 17 patients from controls in colorectal cancer, ulcerative colitis and Crohn's
 18 disease individually, whereas there lacks a systematic analysis to investigate
 19 the exclusive microbial shifts of these enteropathies with similar clinical
 20 symptoms. Our meta-analysis and cross-disease comparisons identified
 21 consistent microbial alterations in each enteropathy, revealed microbial
 22 ecosystems among marker bacteria in distinct states, and demonstrated the

1 necessity and feasibility of metagenome-based multi-disease classifications.
2 To the best of our knowledge, this is the first study that constructed multi-class
3 models in these common intestinal diseases.

4 Background

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

16 IBD and CRC share several symptoms, including rectal bleeding,
17 abdominal pain, diarrhoea, weight loss, and anaemia; furthermore, CRC in
18 young patients has similar ages of onset to IBD (<50 years) (8). In addition,
19 patients with IBD are considered at high risk of developing colorectal cancer,
20 due to duration of inflammation and expansion of lesions. The accumulative
21 risk of CRC in IBD patients is increasing over time (9). The European Crohn's
22 and Colitis Organisation (ECCO) and the American Gastroenterological

1 Association (AGA) recommend that the IBD patients need to strengthen CRC
2 surveillance with colonoscopies. But long-period surveillance doesn't solve the
3 problem because of deficiencies of regular colonoscopies in detecting
4 dysplasia and other high risk factors in the elderly patients (10). It thus can be
5 challenging to accurately separate these diseases in clinical practice,
6 especially in their early stages and/or in younger patients; delay in diagnosis is
7 common and can cause harm, as a recent research has pointed out (8).

8 Recent studies suggested that IBD and CRC are linked with complicated
9 interplay of various components, involving genetics, environmental factors, gut
10 microbiome and immune system (11, 12). So far, a few hundreds of genes
11 have been identified, whose mutations and/or dysregulation of expressions
12 were linked to increase risk of IBD and CRC (13-16). However, genetic factors
13 together can only explain limited proportions of the disease incidences (17-19).
14 Conversely, other factors are believed to be major contributors, especially gut
15 microbes (12, 20-22). The latter, along with metabolites and antibiotics that
16 they produced through digesting nutrients from food, the host and other
17 microbes, could play important roles in modulating host immunity and
18 inflammation, maintaining gastrointestinal equilibrium and resisting alien
19 invaders (23).

20 Fecal microbial dysbiosis in IBD and CRC have been observed, and
21 subsequently utilized to generate predictive models for patient stratification
22 and/or risk evaluation (22, 24-26). For example, IBD patients showed a

1 reduction of taxa from Firmicutes phylum and enrichment of pathogenic
2 species (25, 27). Several studies showed the IBD subtypes, namely CD and
3 UC had distinctive gut microbiota and metabolic profiles, though results differ
4 across studies (28). Gut microbes have been identified as potential markers in
5 distinguishing patients from controls in IBD and CRC individually, as both the
6 increase of pathogens and development of lesions in gut contribute to the
7 dysbiosis through affecting metabolic functions of bacteria (24, 29, 30).

8 However, binary models (i.e. models capable of distinguishing patients of a
9 particular disease from controls) created for a single disease may lead to
10 misdiagnosis, on account of some microbes commonly changed in diseases
11 (31). Furthermore, most models, especially those available for IBD and/or its
12 subtypes, were generated on data from a single population and may not
13 perform well on other populations (25, 27). In addition, though limited by the
14 use of 16S amplicon sequencing data with low resolution, a study across
15 multiple diseases to search for disease-specific markers raised the issue that
16 whether we could distinguish one gut illness from others using solely the gut
17 microbiome data (31).

18 In sum, it is necessary to perform cross-disease comparisons and
19 generate multi-class models capable of distinguishing these common intestinal
20 diseases which can have very similar symptoms and associate with consistent
21 gut microbiome alterations. It is also necessary to perform meta-analysis to
22 account for population-specific biases. Meta-analysis is a method combining

diverse projects and help us avoid biases from individual study (32); 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 (33, 34).

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.

13

14 RESULTS

15 Collection and annotation of thirteen gut metagenomics datasets for common
16 intestinal diseases

17 To determine consistently altered gut microbial features in common intestinal
18 diseases such as UC, CD and CRC as comparing with controls, we conducted
19 a systematic search in public databases (Fig. S1) and collected in total thirteen
20 metagenomic datasets, including three, three and seven datasets for UC, CD
21 and CRC respectively, totaling 763 cases and 632 controls (Table S1).
22 MetaPhlAn2 and HUMAnN2 were used to determine the taxonomic and

1 functional profiles of all samples.

2 Taxonomic analysis revealed that the alpha diversity was not significantly
3 changed in all but one CRC datasets (PRJDB4176) as compared with their
4 respective controls; conversely, alpha diversity was decreased significantly in
5 patients of most CD datasets, while did not show consistent trends in UC
6 patients (Table S1, Wilcoxon rank sum test, P-value < 0.05). Interestingly, we
7 found that the human DNA contents (HDCs), as calculated as the percentage
8 of sequencing reads mapped to the human genome, were significantly higher
9 in patients in all diseases except one UC dataset (PRJEB1220) (Table S1),
10 consistent with our results (35) that HDC could be used as a marker for
11 intestinal diseases; the increased level of HDCs are likely due to that high level
12 of deciduous epithelial and/or blood cells could be found in stools of patients
13 with IBD or CRC resulting from gut injury and quickening cell cycles (21, 24, 29,
14 36).

15

16 Disease-specific and shared taxonomic gut-microbiome markers in CRC, UC 17 and CD

18 We used MaAsLin2, a multivariable analysis tool, on relative abundance of
19 species to adjust the confounding factors, such as body mass index (BMI),
20 gender and age, and identify differential species in each dataset. We then
21 performed meta-analysis on each disease to identify microbes that showed
22 consistent trends in the same disease and referred them as “marker species”

1 accordingly. Consequently, we identified in total 14, 43 and 44 marker species
2 in UC, CD and CRC respectively; among which, 8 (57.1% out of 14), 32
3 (74.4%) and 31 (70%) were unique to the respective diseases. Out of the in
4 total 87 marker species, 14 were found in at least two diseases and no one
5 was common to all diseases (Fig. 1A).

6 For CRC, marker species were mostly disease-enriched, including
7 *Fusobacterium nucleatum*, *Parvimonas micra*, *Gemella morbillorum* and
8 *Peptostreptococcus stomatis*, most of which were reported widely (33, 34).
9 Interestingly, a significant proportion of CRC-enriched marker species were
10 identified significantly in majority of CRC datasets, while most of
11 CRC-depleted marker species were dataset-specific (p-value < 0.05 identified
12 by MaAsLin2, Fig. S2), which was also in accordance with previous studies (33,
13 34).

14 Conversely, CD patients showed a depletion of control-enriched species,
15 including *Roseburia inulinivorans*, *Roseburia hominis*, *Coprococcus catus* and
16 several members of genus *Alistipes*, *Bacteroides*, and *Eubacterium*, which
17 were also consistent with previous studies (25, 27). However, marker species
18 in UC were a mix of both but mainly driven by disease-enriched ones (Fig. 1A),
19 unlike a recent study that primarily showed a decreasing of control-enriched
20 species (27); the discrepancies are likely due to study-specific results (Fig. 1A).
21 Moreover, most of UC- and CD- marker species were identified as significant
22 differential species in at least two datasets, unlike the CRC marker species

1 (p-value < 0.05 identified by MaAsLin2, Fig. S2).

2 Among the shared markers, *Alistipes onderdonkii* and *Ruminococcus*
3 *torques* showed conflicted trends between disease; for example, they both
4 were decreased in CD patients but increased in CRC patients. These results
5 are in fact consistent with previous studies (37-40) and suggest that both
6 overgrowth and loss of certain species represent the disturbance of intestinal
7 environment.

8 Disease-enriched species are often directly linked to pathogenesis and
9 direct targets for disease intervention. We thus first focused on these species
10 and grouped them into disease-specific and common according to their shifts
11 in intestinal diseases. As shown in Fig. 1A, in total 6, 10, 31 and 6 markers
12 were assigned to UC-specific, CD-specific, CRC-specific and
13 common-enriched groups respectively. Their phylogenetic relationships based
14 on the NCBI taxonomic tree revealed their distinct distributions (Fig. S3).
15 Firstly, CRC-specific markers whose taxonomic level were across six phyla,
16 showed more diverse phylogenetically than CD- and UC-specific markers.
17 CD-specific markers were members of two phyla, namely Proteobacteria and
18 Firmicutes, including species within *Veillonella* genus, Enterobacteriaceae
19 family, and Lactobacillales order. The UC-specific markers consist of species
20 from Firmicutes. In addition, UC-specific species are phylogenetically closer to
21 CRC-specific species, likely because UC patients have a little higher risk of
22 CRC (9). Together, we revealed phylogenetic patterns of the marker microbes

1 that could only be revealed through cross-disease analysis.

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

16 We then summed up the prevalence of the disease-enriched species of
17 each sample. As expected, these disease-specific markers were significantly
18 enriched in their respective diseases, as well these common microbes were
19 significantly enriched in all diseases (Fig. 1C). The UC-specific species,
20 though lacked a strong co-occurrence among themselves, had a remarkable
21 prevalence in each disease dataset. The results suggest that with the
22 disease-specific clusters, it would be possible to stratify different diseases

1 using microbial profiles, while the shared enriched species increased
2 difficulties of classification.

3

4 Disease-specific and shared functional markers in CRC, UC and CD

5 Using the same criteria, we identified in total 10, 37 and 39 marker pathways
6 for UC, CD and CRC (Fig. S4); among which, 3 (30% out of 10), 18 (48.6%)
7 and 25 (64.1%) were unique to these diseases respectively. Most of the UC
8 and CD marker pathways were control-enriched and associated with
9 biosynthesis, consistent to previous results (27). For example, pathways for
10 amino acid biosynthesis, such as L-methionine biosynthesis I, and aspartate
11 superpathway were depleted in CD patients, indicating the nutrient transport
12 and uptake (28, 41). Conversely, the gut community in CRC showed distinct
13 characteristics with the decreased capacities of carbohydrate degradation, and
14 increased capacities on amino acid degradation, which were accordance with
15 previous studies (33, 34).

16 To provide an overview on the changed functional capacities of gut
17 microbes, we summarized the metabolic functions as the modules according
18 to their superclasses in MetaCyc database (42). We applied the differential
19 abundance analysis in module level, and found the characteristic functional
20 pattern of IBD and CRC (Fig. 2). As mentioned previously, the module of
21 amino acid degradation was decreased in CD patients, while its trend behaved
22 in the opposite way in CRC patients. In CD patients, the elevated module

1 about nucleoside and nucleotide degradation, which was comprised of
 2 degradation of purine, would induce the gut metabolic stress and involve in the
 3 inflammatory processes (43, 44). As essential pathways for energetic and
 4 biosynthetic demands of cancer cells, the carbohydrate biosynthesis,
 5 fermentation and glycolysis were enhanced in CRC (45, 46). In addition, as
 6 subtypes of IBD, CD and UC had distinct preferences in module level, though
 7 there was a high-degree overlapping among their associated pathways (Fig.
 8 S4).

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

14

15 Disease marker microbes underlie altered metabolic capacities, especially in 16 degradation

17 To check if the altered marker species could underlie the changes in metabolic
 18 capacities, we calculated partial Spearman's rank correlations between marker
 19 species and metabolic pathways and performed meta-analysis to aggregate
 20 coefficients. Interestingly, we found that most of the disease-altered pathways
 21 showed statistically correlations with the marker species; more importantly, we
 22 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 (Fig. 3 & Fig. S5). For example, most of control-enriched species in CD and CRC, including members of genus *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 fermentating general carbohydrates and producing butyrate, which are of anti-inflammatory effects in the gut (47, 48). 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 (49, 50). Similarly, some IBD-depleted species, such as *Alistipes shahii*, *Subdoligranulum spp.*, and *Ruminococcus bromii*, had a negative association with superpathway of purine deoxyribonucleosides degradation, a pathway used as the source of energy (51) (Fig. 3). Of note, the pathways relating to amino acid degradation are positively associated with most of 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

1 studies and diseases revealed differences in metabolism among patients with
2 different diseases, particularly between CRC and IBD.

3 The gut metabolic properties are known to be influenced by food and
4 microbial activities. Recently, researchers also revealed that cells/metabolites
5 derived from the human host, likely due to compromised intestinal barrier (CIB),
6 can also influence the growth of individual bacterium and the gut microbes as a
7 whole (25, 28). CIB could lead to increased HDCs in the gut metagenomics. As
8 expected, we found that HDCs were significantly elevated in cases of all
9 datasets except PRJEB1220 (Table S1). Surprisingly, we found HDCs were
10 also significantly correlated with some CRC-enriched species and half of
11 control-enriched marker species in CD (Fig. S6). *Eubacterium ventriosum*, the
12 control-enriched bacterial marker in CD datasets, was negatively correlated
13 with HDCs in UC and CD ($\rho = -0.32$, $P\text{-value} = 8.16\text{e-}12$ and $\rho = -0.25$,
14 $P\text{-value} = 2.81\text{e-}6$ respectively, Spearman's rank correlation); the correlation
15 was not significant in CRC, most likely due to its low abundances. *E.*
16 *ventriosum* was previously shown to negatively correlate with fundamental
17 components of eukaryotic cell membranes (25). Only three control-enriched
18 species in UC had associations with HDCs (Fig. S6A), while most of
19 UC-depleted pathways correlated with HDCs (Fig. S6B), implying that the
20 metabolic functions had a better response to the intestinal status.

21 Together, our results revealed correlated changes between marker
22 species and metabolic pathways, and suggested that both species and

1 metabolic functions could be driven by the increased human derived contents
2 leaked into the gut due to CIB, consistent with our previous results (35).

3

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

6 Having shown that the alterations in intestinal ecosystems could contribute to
7 gut microbiota dysbiosis, we further explored the inter-relationships among the
8 marker species within each physical condition. As ecologically important
9 patterns, coexistence relationships within a biological community could reflect
10 interplays between organisms and ecological roles of individual members.
11 Applying co-occurrences analyses to gut microbes could help us compare
12 coexistence patterns from different intestinal states, identify key species
13 important to human health and provide an insight into maintenances of gut
14 microbial ecosystems (52, 53).

15 We thus constructed inter-species networks using the disease marker
16 species separately for cases and controls for each disease based on pairwise
17 correlations of the species abundances. We used SparCC, a correlation
18 method for microbiome data, to calculate the correlation coefficients among
19 species to perform meta-analysis. We found that species in the cases were
20 connected more often than they were in the controls of respective disease
21 datasets (Fig. 4). For example, we found 172 positive pairs and 45 negative
22 pairs of correlated marker species (fdr in meta-analysis < 0.05) in CRC,

1 increased from 138 and 36 in the controls relatively (Fig. 4A & B, Fig. S7A & B).
 2 Similarly, we found much more positive and negative correlations among
 3 markers in cases than in controls in both CD and UC (Fig. 4A & B, Fig. S7C-F).
 4 These results were coincide with a previous study, which identified that the
 5 CRC patients networks contained more links among nodes than control
 6 networks, and the negative correlations declined when CRC patients under the
 7 chemotherapy (54).

8 In view of the difficulties to compare the networks as a whole, we used the
 9 mcode implemented in Cytoscape to detect modules, which are regarded as
 10 the highly interconnected clusters in a network and often used to gain
 11 biological insights from networks. In CRC datasets, we found that the modules
 12 from cases (named as 'CRC-Case1') had more members and was
 13 interconnected tighter as compared with those from controls (named as
 14 'CRC-CTR1' and 'CRC-CTR2'), (Fig. 4C). The species belonging to the same
 15 genus were associated more closely with each other, probably owing to their
 16 similar metabolic properties. Our network-derived modules could reveal
 17 previous known positive interactions. For example, *F. nucleatum*, a widely
 18 studied oral-associated anaerobe known to co-aggregate with other
 19 anaerobes to form biofilm and involve in the intestinal tumorigenesis (55, 56),
 20 showed positive correlations with *P. micra* in our case module (Fig. 4C) (57).
 21 Besides, though it lacks experimental evidences for our identified negative
 22 associations underlying CRC-enriched microbes, there were plenty of

1 investigations about the competitive relationship among taxa during growth of
2 biofilm (57-60). For example, *Porphyromonas gingivalis*, another known
3 biofilm-forming partner of *F. nucleatum*, showed a negative correlation with *P.*
4 *micra* (57). Thus, our results found the novel relations between CRC-enriched
5 microbes and remained to be confirmed in further experiments.

6 Similar with CRC modules, modules from CD datasets also displayed the
7 tighter relationships within the species at same taxonomic level, including
8 members of genus *Klebsiella*, *Veillonella* and *Alistipes* (Fig. 4D). A previous
9 study found that *Klebsiella* correlated positively with fecal calprotectin (FCP),
10 an inflammatory marker for IBD, whereas *Ruminococcus* correlated negatively
11 with FCP (53). In UC datasets, only the module from UC patients' network
12 (named as 'UC-Case1') was recognized (Fig. 4E). The strong correlation
13 between *G. morbillorum* and *P. stomatis* was shown in UC module and CRC
14 modules despite the sources of data, indicating the co-aggregation between
15 them.

16 We also identified potential hub species in the networks using eigenvector
17 centrality scores (ECSs) and betweenness centrality scores (BCSs) (Fig. S8).
18 ECS served as assessment of node influence in a weighted network,
19 measures the importance of the given node considering not only its
20 connections with others, but also the connections of its related nodes. BCS
21 was used to evaluate the transmission capacity of species. In CD datasets, we
22 found that *Alistipes putredinis*, *A. shahii* and three CD-enriched *Veillonella*

1 were the pivotal species in controls; while in CD patients, two CD-enriched
2 Enterobacteriaceae, *Citrobacter spp.* and *Klebsiella spp.* took leading roles
3 (Fig. S8C & D), both of which could deliver virulence proteins into host cells for
4 assisting themselves against host immune system and infect mucosa so that
5 thrive in gut (61-63). Nevertheless, health-related species were also the hub
6 bacteria in patients with CD. These results support the view that we should be
7 cautious to use antibiotic in CD therapy, which may disrupt fragile connections
8 among species exhausting in cases, and cause some bacteria failure of
9 recovery (64). As expected, CRC-enriched microbes, such as most of species
10 from genus *Clostridium*, were at the center in both control network and patient
11 network of CRC datasets (Fig. S8A & B). Although the top nodes with high
12 BCSs were CRC-enriched species in both CRC control- network and CRC
13 case- network, their niches changed, suggesting that the CRC-enriched
14 microbes were vital in the dynamic network. In UC datasets, nodes with
15 highest ECSs in controls network were UC-depleted species, while in the case
16 network were mainly UC-enriched species (Fig. S8E & F). We found that the
17 crucial nodes were also the members of corresponding module, validating
18 representativeness of the modules in network.

19 Together, our inter-species network analysis revealed that marker species
20 were more closely connected in diseased conditions; we speculated that due
21 to oxidative stress and increased permeability of intestinal barrier, the gut
22 ecosystem under diseased states may represent more stressful conditions, in

1 which the growth of all microbes, especially the marker species, would be
2 under stronger constraints and selection (11, 65). In addition, the hub species
3 that were at the center positions and connect more with others, are more likely
4 to be targets for disease treatment.

5

6 Multi-class machine-learning models for disease stratification using gut 7 microbial and metabolic markers

8 We next built multi-class models capable of distinguishing multiple diseases
9 using the microbial and pathway profiles as well as those of the identified
10 markers.

11 We first built four-class models with multiple types of features mentioned
12 above using 10-time 10 folds cross-validation (see Materials and Methods). A
13 model based on combined taxonomic and metabolic profiles (all features)
14 reached a highest accuracy with true positive rate (TPR) of 0.8 (Fig. S9A & B).
15 We achieved significantly better classification performance on UC and CD
16 than a recent study, with TPRs of 0.81 and 0.91 for UC and CD samples
17 respectively in this model (Fig. S9B), as compared to 0.51 and 0.67 in the Fig.
18 6 of literature (25) where the model was also trained on combined profile.
19 Models built on either taxonomic (TPR as 0.69) or functional profiles (0.74)
20 showed decreased performance (Fig. S9A). The classifier based on combined
21 markers also performed well with an overall accuracy of 0.75 (Fig. 5B). The
22 classification errors were mostly contributed by UC samples, which associated

1 with the least number of disease markers and the least proportion of unique
 2 ones (Fig. 1A & Fig. S4). Furthermore, UC shared majority of its markers with
 3 CD, but not *vice versa*; consequently, a significant proportion of UC samples
 4 were classified as CD, but only a few of CD samples were misclassified as UC
 5 (Fig. 5B).

6 Since UC and CD are sub-types of IBD, we thus combined their samples
 7 into the IBD group and built three-class models (i.e., IBD, CRC and controls,
 8 Fig. 5C & Fig. S9A); as shown in Fig. 5C, we achieved much better
 9 classification rate with an overall TPR of 0.81 using also the combined markers.
 10 90% of the IBD samples were classified correctly; the remaining 10% were
 11 misclassified as control, but none was classified as CRC.

12 We also built two additional models, including a “case-control” model to
 13 distinguish cases from controls (also referred to as “binary” model), and a
 14 “cases” model to assign cases to distinct diseased states (also referred to as
 15 “cases” model). The cases model was particularly important and clinically
 16 relevant due to the clinical overlap in presentation of these diseases, as well as
 17 the risk of IBD patients to eventually develop CRC. We evaluated the
 18 performance for the binary models on the 632 controls and 763 cases and
 19 found that all models performed well, with the area under receiver operating
 20 characteristic curves (AUROCs) ranging from 0.80 to 0.88. Notably,
 21 taxonomic-based models in generally performed better than the
 22 metabolic-based models, as the model with the combined profile attained a

1 highest accuracy (Fig. 5A). Surprisingly but expected, models using only the
 2 marker species/pathways performed comparable to those that used all
 3 species/pathways, especially the combined-based model, suggesting that the
 4 much-shortened list of markers are of practical and clinical value. For the
 5 “cases” models, the classifier based on combined profile achieved an accuracy
 6 of 0.98, while the accuracy of classifier based on combined markers only
 7 achieved 0.9. The metabolic-based model still performed better than
 8 taxonomic-based model, indicating the functions of microbes reflected the gut
 9 status better than species distribution (Fig. 5D, Fig. S9A). We noticed
 10 significant differences among diseases in terms of TPRs in the “cases” models.
 11 In particular, we achieved high accuracies for CRC and CD (TPR as 99% and
 12 93% respectively), as compared with the relatively low TPR for UC (68%). The
 13 latter was likely due to the fact that UC shared most of its markers with CD and
 14 had only few unique markers; consequently, most of the misclassified UC
 15 cases were predicted as CD (Fig. 1A & Fig. S4).

16 To evaluate if the performances were not biased by a single dataset, we
 17 applied leave-one-dataset-out (LODO) analysis, which left one dataset as the
 18 testing data, and utilized the remained datasets to train the random forest
 19 models. The LODO models based on combined profile and combined markers
 20 (referred to as “Four-class All” and “Four-class Dif” relatively) to distinguish
 21 cases with the different disease and controls, achieved an average accuracy of
 22 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 of each dataset were similar, indicating there was no bias across datasets. Moreover, except the models for classifying patients with different disease, the models trained on the combined markers performed better than the corresponding models trained on combined profile.

These results suggest the classifiers based on combined markers could achieve similar accuracies with those based on combined profile, indicating the clinical feasibility of the microbial markers. Besides, 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.

DISCUSSION

In this study, we collected fecal metagenomics datasets for three common intestinal diseases, namely CRC, CD and UC, totaling 11 projects, 13 datasets, 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 common clusters

1 according to whether or not the member species were unique to a certain
2 disease, and revealed their distinct phylogenetic relationships: for example,
3 CRC- specific species are more diverse phylogenetically than UC- and
4 CD-specific markers. Strikingly, UC- and CRC- specific species are
5 phylogenetically closer to each other than to those of CD (Fig. S3), in part due
6 to the fact that UC-patients are at higher risk of developing CRC (9).

7 We then characterized the marker pathways. We first revealed that each
8 disease formed their exclusive module profiles, that the CRC patients had an
9 elevated trend in amino acid degradation while the CD patients behaved in an
10 opposite way. We then explored that almost all marker pathways correlated
11 significantly with marker species (Fig. 3 & Fig. S5); additionally, clustered
12 marker species tended to correlate significantly with the same sets of
13 pathways. These results were not unexpected since marker species that are
14 closer phylogenetically tend to have similar metabolic capacities. We then
15 noticed strong correlations between a significant proportion of marker species
16 and HDCs. HDC has been shown to be significantly increased in many
17 intestinal diseases, and could be used as an indicator for the extent of leaky
18 gut caused by CIB (Table S1). The elevated HDCs in all datasets may signify
19 significant changes in physio-metabolic properties of the local gut environment
20 due to leaked human derived contents under diseased states. Our results thus
21 suggested that human derived contents due to CIB could have stronger impact
22 on gut microbiota than we have previously anticipated. Finally, by considering

1 the gut microbiota as an ecosystem, we revealed that marker species showed
2 increased connectivity in diseases as compared with the respective controls,
3 and control-enriched species together with pathogens played important roles
4 in ecological network of CD patients. Thus we speculate that the diseased gut
5 may represent a more stressful environment due to the physio-metabolic
6 changes including oxidative stress and/or bleeding; the inhabitant microbes
7 are thus under stronger selection, and show either more cooperation (positive
8 correlation) or competition (negative correlation). And our results support the
9 view that we should be cautious to use antibiotics in therapies for CD patients.

10 Utilizing the identified marker species and pathways, we obtained four
11 high-performance models for disease identification and patient stratification.
12 The first “four-class” model could separate samples into controls or individual
13 diseases (Fig. 5B), with an overall TPR of 0.75. UC has the lowest TPR (0.37,
14 Fig. 5B) in this model; however, most of the wrongly classified samples went to
15 CD, consistent with previous efforts (25) and the fact that UC had very few
16 unique markers and shared most of its marker with CD (Fig. 1A and Fig. S4).
17 Regardless, it represents one of the best models that could classify IBD
18 subtypes with TPR as 0.81 and 0.91 separately in UC and CD samples, while
19 in a previous study TPR from the cross-validation model built on the
20 abundance of metabolites and species only achieved 0.49 and 0.66 in UC and
21 CD respectively (25). We also built three additional models, including a “binary”
22 model to distinguish cases from controls, a “three-class” model to consider the

1 IBD as a whole and distinguish patients with cancer or inflammation from
2 controls, and a “cases” model to assign cases to distinct diseased states. In
3 our opinion, both are relevant in clinical applications. For example, the binary
4 model, with AUROC of 0.87, can inform the subjects for further clinical
5 inspections such as colonoscopy, while the “three-class” model, with an overall
6 TPR of 0.81, can evaluate the patients for potential IBD and CRC risks. The
7 “cases” model with a high accuracy of 0.9 was worth watching, due to the
8 clinical overlap in symptoms of these intestinal diseases, as well as the risk of
9 IBD patients to eventually develop CRC.

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

19

20 MATERIALS AND METHODS

21 Data collection and pre-processing

22 We obtained in total 175 records by searching public metagenomic databases

1 including NCBI PubMed (66) and GMrepo (67) using key words such as
 2 metagenomics and relevant disease names (see Fig. S1 for details). We
 3 aimed to collect metagenomics sequencing data with high-resolution for better
 4 understanding the functions of microbes. After filtering out the duplicates, 16s
 5 rRNA sequencing data and the metagenomics data without detailed metadata
 6 or not meeting minimum samples requirements, we selected in total thirteen
 7 metagenomics datasets, including three, three, and seven datasets for CD, UC
 8 and CRC respectively. Please consult Fig. S1 for the selection procedure and
 9 results, and Table S1 for the thirteen datasets.

10 Raw sequencing reads were retrieved from European Nucleotide Archive
 11 (ENA) (68) under the following identifiers: PRJEB6070 (22), PRJEB27928 (33),
 12 PRJEB12449 (69), PRJEB10878 (26), PRJEB7774 (39), PRJDB4176 (33),
 13 cohort 1 of PRJNA447983 (33), SRP057027 (24), PRJEB1220 (70),
 14 PRJNA400072 (25) and PRJNA389280 (71); sample metadata were also
 15 downloaded from ENA. For projects containing samples resulting from
 16 longitudinal surveys, i.e., participants were sampled multiple times over
 17 extended periods of time and/or during treatment/intervention, including
 18 SRP057027, PRJEB1220 and PRJNA389280, we selected the first time-point
 19 from each participant to avoid false positive in following analysis. In total, we
 20 obtained in total 632 non-disease controls and 763 patients for the following
 21 meta-analysis, including 354, 177 and 232 samples of CRC, UC and CD
 22 (Table S1).

1 Taxonomic and functional profiling of metagenomics data

2 To keep only the high quality data, low quality reads and adapters were firstly
 3 removed via Trimmomatic (version 0.35) using the Truseq3 adapter files
 4 (TruSeq3-PE.fa for paired-end data and TruSeq3-SE.fa for single-end data)
 5 and a MINLEN cutoff of 50 (72). The remaining “clean” reads were then
 6 mapped to the human reference genome (hg19) using bowtie2 (version 2.3.4.3)
 7 (73) with default settings to identify and remove human reads. The identified
 8 human reads were also used to compute HDC for each sample as the
 9 percentage of mapped reads out of total clean reads, which have been shown
 10 to be a marker for intestinal barrier dysfunction and correlate with the marker
 11 species of several intestinal diseases (35). For samples that were sequenced
 12 multiple times (e.g., for the purpose of increasing sequencing depths), the
 13 resulting multiple sequencing files were merged before further analysis. The
 14 merged and clean non-human reads were then quantified in taxonomic and
 15 functional levels using MetaPhlAn2 mapping to mpa_v20_m200 database and
 16 HUMAnN2 mapping to ChocoPhlAn database and full UniRef90 database (74,
 17 75).

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

1 analyzed.

2 Controlling for confounding factors and identification of marker species and 3 pathways

4 Within-project confounding factors, i.e., those showed significant differences
5 between phenotype groups in a dataset were firstly identified using Wilcoxon
6 rank sum test or chi-squared test on per-dataset basis. Then, the identified
7 confounding factors (see Table S1 for the results) in differential analysis were
8 controlled for using MaAsLin2 package in R ver 4.0.0, a multivariable analysis
9 tool to adjust the covariates and identify association effects of the species and
10 pathways to disease in each dataset. Accounting for the heterogeneity
11 between datasets, we performed meta-analysis to aggregate the association
12 effects via MMUPHin package in R ver 4.0.0, and identify the final “maker”
13 species and pathways. Here an adjusted p-value < 0.05 from meta-analysis
14 was used as the cutoff for the makers.

15 Clustering of disease-enriched species and their prevalence in the three 16 diseases

17 Consistently disease-enriched marker species (i.e., those were marker
18 species in at least two datasets of the same disease) were grouped into
19 disease-specific or common to multiple diseases according to their association
20 with the diseases (Fig. 1A). To observe the prevalence of the clusters in the
21 overall patients and a unique disease (i.e., CRC, CD and UC), their prevalence
22 in the diseased samples were first calculated: for each selected marker

1 species, its 95% percentile abundance in all controls was used as a cutoff to
 2 defined its presence (“1”, i.e., its relative abundance in the sample was higher
 3 than the 95% quantile relative abundance of all control samples) or absence
 4 (“0”). By doing so, we obtained a binarized matrix with each row representing a
 5 disease-enriched marker microbe, and each column representing a patient.
 6 The prevalence matrix from all patients was used to calculate the Jaccard
 7 distances among the species using the diversity function of vegan package.
 8 We compared the inner Jaccard similarities among the clusters of
 9 disease-enriched marker species using Wilcoxon rank sum test (for pairwise
 10 comparisons) and Kruskal-Wallis rank sum test (for multi-group comparisons).
 11 The prevalence of each cluster between patients and controls in each disease
 12 were also compared using the Cochran-Mantel-Haenszel test with “dataset” as
 13 the blocked object by cmh_test function of coin package.

14 **Phylogenetic relationship of disease-enriched marker species**

15 To show the phylogenetic relationships among the disease-enriched marker
 16 species, a phylogenetic tree was generated based on their NCBI taxonomy
 17 using an online tool, phyloT (<https://phylot.biobyte.de/>), setting internal nodes
 18 collapsed and polytomy as no. The tree file then was visualized using Evolview
 19 ver3, a webserver for annotation and management phylogenetic trees (76).
 20 The nodes were colored depending on their corresponding clusters identified
 21 in previous section. The last common ancestors (LCAs) were determined
 22 according to the NCBI taxonomy of the species in corresponding branch.

1 Identification of HDC-correlated features

2 For each dataset, Spearman's rank correlation was used to identify
3 HDC-related microbial features (e.g., species and functions) using a p-value
4 cutoff of 0.05. Features that maintained a significant positive or negative
5 relationship with HDC in at least two datasets of a disease were identified as
6 HDC-related features.

7 Functional profile of metabolic modules

8 According to the category of MetaCyc database, we concluded the microbial
9 functions into their corresponding superclasses as metabolic modules (42).
10 The expression of each metabolic module was summarized as the average
11 logarithm relative abundance of its contained functions. Setting the quantiles
12 from 0.1 to 0.9 and the increment as 0.1, we calculated the generalized fold
13 changes of modules between the controls and cases, and performed the
14 Wilcoxon rank sum test with the “datasets” as the blocked object to evaluate
15 the differences.

16 Microbial ecosystem analyses using species-species correlations

17 To characterize the relationships among the marker species and the resulting
18 interaction networks, SparCC, a sparse correlation method for compositional
19 data (77) was used to identify correlations among marker species. SparCC
20 was previously shown to be able to reduce the high false positive by
21 Spearman's rank correlation in metagenomics data. The tool requires read
22 counts as input, therefore we multiplied the relative abundances of the species

1 to the number of reads mapping to mpa_v20_m200 database, and got the
 2 microbial counts of each sample. For each dataset, species-species
 3 correlations were calculated for control and case samples separately. By
 4 setting both the iteration number and simulation as 100 and threshold of
 5 correlation strength as 0.05, SparCC generated the correlation matrices of the
 6 real data and 100 simulated datasets. The pseudo P-values were assessed as
 7 the proportion of simulated datasets with a correlation value at least as
 8 extreme as that calculated from the real data. After filtering correlations with
 9 $P\text{-value} < 0.05$, we performed meta-analysis to aggregate correlation
 10 coefficients for each disease type via random-effects model, which
 11 summarizes overall correlation based on Fisher's z transformation with
 12 metacor function (32, 78). The summarized correlations with adjusted p-values
 13 < 0.05 in meta-analysis were used to construct networks. The networks were
 14 analyzed in Cytoscape (79) to identify modules using mcode with default
 15 parameters. We then evaluated eigenvector centrality and betweenness
 16 centrality of networks using correlation strength as weight, and visualized
 17 networks with igraph package in R. The size of node indicated prevalence of
 18 the bacteria in counterpart samples. Positive and negative correlation
 19 coefficients as strength of edges were painted gray and red separately.

20 **Correlating functional profiles with species**

21 To identify species underlying functional changes in the metagenomic data,
 22 correlations between relative abundance values of marker species and marker

1 metabolic pathways in each dataset were computed using partial Spearman's
2 rank correlation to adjust the identified covariates. The relative abundances
3 were log-transformed; to avoid Inf values, pseudo values 1e-06 and 1e-09
4 respectively were added to taxonomic and functional abundances before
5 log-transformation. The resulting correlations with P-value <0.05 were retained
6 to perform meta-analysis and get the overall correlation coefficients with
7 metacor function.

8 Random forest classifiers and cross validation

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

17 For the overall cross-validation, we pooled datasets into one, then applied
18 logarithm transformation and standardization to the taxonomic and functional
19 abundance profiles. The data were split into training set and testing set
20 repeatedly 10 times. All models trained on the training set were applied to
21 corresponding testing set and the prediction scores were averaged. For binary
22 classifiers, i.e., classifiers that attempt to classify samples into two distinct

1 groups: diseased or control, the AUROC values were used to evaluate their
 2 performance. For the multi-class classifiers, i.e., classifiers that attempt
 3 classify samples into multiple distinct groups such as CRC, UC, CD and
 4 control, the detailed predicting results were shown as confusion matrixes and
 5 the performances as overall accuracies. We also built models based on the
 6 identified microbial markers to test if the performances of models were
 7 improved.

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

16 Other statistical tests

17 We calculated alpha diversity of each dataset with the diversity function of
 18 vegan package. Two-sided Wilcoxon rank sum test was used to compare two
 19 sets of numeric data; the wilcox_test function implemented in the coin package
 20 was used to block with the factor “dataset”.

21 Data availability

22 The datasets generated and R codes during this study are available at

1 <https://github.com/whchenlab/2019-puzi-multi-gut-disease-classifier>.

2 Correspondence and requests for materials should be addressed to W.H.C.

3 and X.M.Z...

4

5 **ACKNOWLEDGMENTS**

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

7

8 This work was partly supported by National Key Research and

9 Development Program of China [2019YFA0905600 to W.H.C.], National

10 Natural Science Foundation of China [61932008, 61772368, 61572363 to

11 X.M.Z], National Key R&D Program of China [2018YFC0910500 to X.M.Z],

12 Natural Science Foundation of Shanghai [17ZR1445600 to X.M.Z], Shanghai

13 Municipal Science and Technology Major Project [2018SHZDZX01 to X.M.Z]

14 and ZJLab. The funders had no role in study design, data collection and

15 interpretation, or the decision to submit the work for publication.

16 W.H.C. and X.M.Z. designed the study. P.J. and S.W. collected and

17 analyzed the data, Q.L. coordinated the data downloads and analysis. W.H.C.,

18 and P.J. wrote the manuscript with all authors contributing to the writing and

19 providing feedbacks. All authors read and approved the final version of the

20 manuscript.

21 We declare no competing interests.

22

REFERENCES

1. 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.
2. 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* 8:22529-42.
3. 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* doi:10.1136/gutjnl-2018-317592.
4. Araghi M, Soerjomataram I, Bardot A, Ferlay J, Cabasag CJ, Morrison DS, De P, Tervonen H, Walsh PM, Bucher 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* doi:10.1016/S2468-1253(19)30147-5.
5. M'Koma AE. 2013. Inflammatory bowel disease: an expanding global health problem. *Clin Med Insights Gastroenterol* 6:33-47.
6. Ananthakrishnan AN. 2015. Epidemiology and risk factors for IBD. *Nat Rev Gastroenterol Hepatol* 12:205-17.

- 1 7. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI,
2 Panaccione R, Ghosh S, Wu JCY, Chan FKL, Sung JJY, Kaplan GG.
3 2018. Worldwide incidence and prevalence of inflammatory bowel
4 disease in the 21st century: a systematic review of population-based
5 studies. *Lancet* 390:2769-2778.
- 6 8. Stapley SA, Rubin GP, Alsina D, Shephard EA, Rutter MD, Hamilton
7 WT. 2017. Clinical features of bowel disease in patients aged <50 years
8 in primary care: a large case-control study. *Br J Gen Pract*
9 67:e336-e344.
- 10 9. Farraye FA, Odze RD, Eaden J, Itzkowitz SH. 2010. AGA Technical
11 Review on the Diagnosis and Management of Colorectal Neoplasia in
12 Inflammatory Bowel Disease. *Gastroenterology* 138:746-774.e4.
- 13 10. Abdalla M, Herfarth H. 2018. Rethinking Colorectal Cancer Screening in
14 IBD, Is It Time to Revisit the Guidelines? *Journal of Crohn's and Colitis*
15 12:757-759.
- 16 11. Ni J, Wu GD, Albenberg L, Tomov VT. 2017. Gut microbiota and IBD:
17 causation or correlation? *Nat Rev Gastroenterol Hepatol* 14:573-584.
- 18 12. O'Keefe SJD. 2016. Diet, microorganisms and their metabolites, and
19 colon cancer. *Nature Reviews Gastroenterology & Hepatology*
20 13:691.
- 21 13. Khor B, Gardet A, Xavier RJ. 2011. Genetics and pathogenesis of
22 inflammatory bowel disease. *Nature* 474:307-317.

14. 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 JJY, Malekzadeh R, Westra H-J, Yamazaki K, Yang S-K, Barrett JC, Franke A, Alizadeh BZ, Parkes M, B K T, Daly MJ, Kubo M, Anderson CA, Weersma RK, International Multiple Sclerosis Genetics C, International IBDGC. 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics* 47:979-986.
15. 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.
16. 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.
17. 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.
18. Foulkes WD. 2008. Inherited susceptibility to common cancers. *N Engl J Med* 359:2143-53.

19. 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.
20. Lane ER, Zisman TL, Suskind DL. 2017. The microbiota in inflammatory bowel disease: current and therapeutic insights. *J Inflamm Res* 10:63-73.
21. Dickinson BT, Kisiel J, Ahlquist DA, Grady WM. 2015. Molecular markers for colorectal cancer screening. *Gut* 64:1485-94.
22. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Bohm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Tran Van Nhieu J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, von Knebel Doeberitz M, Sobhani I, Bork P. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 10:766.
23. Rooks MG, Garrett WS. 2016. Gut microbiota, metabolites and host immunity. *Nat Rev Immunol* 16:341-52.
24. 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,

- 1 Bushman FD. 2015. Inflammation, Antibiotics, and Diet as
- 2 Environmental Stressors of the Gut Microbiome in Pediatric Crohn's
- 3 Disease. *Cell host & microbe* 18:489-500.
- 4 25. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ,
- 5 Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, Sauk JS, Wilson
- 6 RG, Stevens BW, Scott JM, Pierce K, Deik AA, Bullock K, Imhann F,
- 7 Porter JA, Zhernakova A, Fu J, Weersma RK, Wijmenga C, Clish CB,
- 8 Vlamakis H, Huttenhower C, Xavier RJ. 2019. Gut microbiome structure
- 9 and metabolic activity in inflammatory bowel disease. *Nat Microbiol*
- 10 4:293-305.
- 11 26. Yu J, Feng Q, Wong SH, Zhang D, Liang Qy, Qin Y, Tang L, Zhao H,
- 12 Stenvang J, Li Y, Wang X, Xu X, Chen N, Wu WKK, Al-Aama J, Nielsen
- 13 HJ, Kiellerich P, Jensen BAH, Yau TO, Lan Z, Jia H, Li J, Xiao L, Lam
- 14 TYT, Ng SC, Cheng AS-L, Wong VW-S, Chan FKL, Xu X, Yang H,
- 15 Madsen L, Datz C, Tilg H, Wang J, Brnner N, Kristiansen K,
- 16 Arumugam M, Sung JJ-Y, Wang J. 2017. Metagenomic analysis of
- 17 faecal microbiome as a tool towards targeted non-invasive biomarkers
- 18 for colorectal cancer. *Gut* 66:70.
- 19 27. Vich Vila A, Imhann F, Collij V, Jankipersadsing SA, Gurry T, Mujagic Z,
- 20 Kurilshikov A, Bonder MJ, Jiang X, Tigchelaar EF, Dekens J, Peters V,
- 21 Voskuil MD, Visschedijk MC, van Dullemen HM, Keszthelyi D, Swertz
- 22 MA, Franke L, Alberts R, Festen EAM, Dijkstra G, Masclee AAM, Hofker

- 1 MH, Xavier RJ, Alm EJ, Fu J, Wijmenga C, Jonkers D, Zhernakova A,
2 Weersma RK. 2018. Gut microbiota composition and functional
3 changes in inflammatory bowel disease and irritable bowel syndrome.
4 Sci Transl Med 10.
- 5 28. Schirmer M, Garner A, Vlamakis H, Xavier RJ. 2019. Microbial genes
6 and pathways in inflammatory bowel disease. Nature Reviews
7 Microbiology 17:497-511.
- 8 29. Kim M, Ashida H, Ogawa M, Yoshikawa Y, Mimuro H, Sasakawa C.
9 2010. Bacterial interactions with the host epithelium. Cell Host Microbe
10 8:20-35.
- 11 30. Zeng MY, Inohara N, Nuñez G. 2017. Mechanisms of
12 inflammation-driven bacterial dysbiosis in the gut. Mucosal Immunology
13 10:18-26.
- 14 31. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017.
15 Meta-analysis of gut microbiome studies identifies disease-specific and
16 shared responses. Nature Communications 8:1784.
- 17 32. Hunter J, Schmidt F. 1990. Dichotomization of Continuous Variables:
18 The Implications for Meta-Analysis. Journal of Applied Psychology
19 75:334-349.
- 20 33. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M,
21 Beghini F, Manara S, Karcher N, Pozzi C, Gandini S, Serrano D, Tarallo
22 S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma

- 1 H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P,
2 Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F,
3 Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L,
4 Naccarati A, Segata N. 2019. Metagenomic analysis of colorectal
5 cancer datasets identifies cross-cohort microbial diagnostic signatures
6 and a link with choline degradation. Nat Med 25:667-678.
- 7 34. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS,
8 Voigt AY, Palleja A, Ponnudurai R, Sunagawa S, Coelho LP,
9 Schrotz-King P, Vogtmann E, Habermann N, Niméus E, Thomas AM,
10 Manghi P, Gandini S, Serrano D, Mizutani S, Shiroma H, Shiba S,
11 Shibata T, Yachida S, Yamada T, Waldron L, Naccarati A, Segata N,
12 Sinha R, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G. 2019.
13 Meta-analysis of fecal metagenomes reveals global microbial
14 signatures that are specific for colorectal cancer. Nature Medicine
15 25:679-689.
- 16 35. Jiang P, Lai S, Wu S, Zhao X-M, Chen W-H. 2020. Host DNA contents
17 in fecal metagenomics as a biomarker for intestinal diseases and
18 effective treatment. BMC Genomics 21:348.
- 19 36. Vincent C, Mehrotra S, Loo VG, Dewar K, Manges AR. 2015. Excretion
20 of Host DNA in Feces Is Associated with Risk of Clostridium difficile
21 Infection. J Immunol Res 2015:246203.
- 22 37. Alipour M, Zaidi D, Valcheva R, Jovel J, Martínez I, Sergi C, Walter J,

1 Mason AL, Wong GK-S, Dieleman LA, Carroll MW, Huynh HQ, Wine E.
2 2016. Mucosal Barrier Depletion and Loss of Bacterial Diversity are
3 Primary Abnormalities in Paediatric Ulcerative Colitis. *Journal of*
4 *Crohn's and Colitis* 10:462-471.

5 38. Takahashi K, Nishida A, Fujimoto T, Fujii M, Shioya M, Imaeda H,
6 Inatomi O, Bamba S, Andoh A, Sugimoto M. 2016. Reduced
7 Abundance of Butyrate-Producing Bacteria Species in the Fecal
8 Microbial Community in Crohn's Disease. *Digestion* 93:59-65.

9 39. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H,
10 Xu X, Jie Z, Su L, Li X, Li X, Li J, Xiao L, Huber-Schönauer U,
11 Niederseer D, Xu X, Al-Aama JY, Yang H, Wang J, Kristiansen K,
12 Arumugam M, Tilg H, Datz C, Wang J. 2015. Gut microbiome
13 development along the colorectal adenoma–carcinoma sequence.
14 *Nature Communications* 6:6528.

15 40. Zhang Y, Yu X, Yu E, Wang N, Cai Q, Shuai Q, Yan F, Jiang L, Wang H,
16 Liu J, Chen Y, Li Z, Jiang Q. 2018. Changes in gut microbiota and
17 plasma inflammatory factors across the stages of colorectal
18 tumorigenesis: a case-control study. *BMC Microbiology* 18:92.

19 41. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV,
20 Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J,
21 Sands BE, Xavier RJ, Huttenhower C. 2012. Dysfunction of the
22 intestinal microbiome in inflammatory bowel disease and treatment.

1 Genome Biology 13:R79.

2 42. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford
3 PE, Ong WK, Paley S, Subhraveti P, Karp PD. 2020. The MetaCyc
4 database of metabolic pathways and enzymes - a 2019 update. Nucleic
5 Acids Research 48:D445-D453.

6 43. Mars RAT, Yang Y, Ward T, Houtti M, Priya S, Lekatz HR, Tang X, Sun
7 Z, Kalari KR, Korem T, Bhattarai Y, Zheng T, Bar N, Frost G, Johnson
8 AJ, van Treuren W, Han S, Ordog T, Grover M, Sonnenburg J, D'Amato
9 M, Camilleri M, Elinav E, Segal E, Blekhman R, Farrugia G, Swann JR,
10 Knights D, Kashyap PC. 2020. Longitudinal Multi-omics Reveals
11 Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. Cell
12 183:1137-1140.

13 44. Crittenden S, Cheyne A, Adams A, Forster T, Robb CT, Felton J, Ho GT,
14 Ruckerl D, Rossi AG, Anderton SM, Ghazal P, Satsangi J, Howie SE,
15 Yao C. 2018. Purine metabolism controls innate lymphoid cell function
16 and protects against intestinal injury. Immunol Cell Biol 96:1049-1059.

17 45. La Vecchia S, Sebastián C. 2020. Metabolic pathways regulating
18 colorectal cancer initiation and progression. Seminars in Cell &
19 Developmental Biology 98:63-70.

20 46. Oliphant K, Allen-Vercoe E. 2019. Macronutrient metabolism by the
21 human gut microbiome: major fermentation by-products and their
22 impact on host health. Microbiome 7:91.

- 1 47. Louis P, Hold GL, Flint HJ. 2014. The gut microbiota, bacterial
2 metabolites and colorectal cancer. *Nature Reviews Microbiology*
3 12:661-672.
- 4 48. Louis P, Flint HJ. 2017. Formation of propionate and butyrate by the
5 human colonic microbiota. *Environmental Microbiology* 19:29-41.
- 6 49. Colvin H, Nishida N, Konno M, Haraguchi N, Takahashi H, Nishimura J,
7 Hata T, Kawamoto K, Asai A, Tsunekuni K, Koseki J, Mizushima T,
8 Satoh T, Doki Y, Mori M, Ishii H. 2016. Oncometabolite
9 D-2-Hydroxyglurate Directly Induces Epithelial-Mesenchymal Transition
10 and is Associated with Distant Metastasis in Colorectal Cancer.
11 *Scientific Reports* 6:36289.
- 12 50. Han J, Jackson D, Holm J, Turner K, Ashcraft P, Wang X, Cook B,
13 Arning E, Genta RM, Venuprasad K, Souza RF, Sweetman L, Theiss
14 AL. 2018. Elevated d-2-hydroxyglutarate during colitis drives
15 progression to colorectal cancer. *Proc Natl Acad Sci U S A*
16 115:1057-1062.
- 17 51. Vich Vila A, Collij V, Sanna S, Sinha T, Imhann F, Bourgonje AR,
18 Mujagic Z, Jonkers DMAE, Masclee AAM, Fu J, Kurilshikov A,
19 Wijmenga C, Zhernakova A, Weersma RK. 2020. Impact of commonly
20 used drugs on the composition and metabolic function of the gut
21 microbiota. *Nature Communications* 11:362.
- 22 52. Williams RJ, Howe A, Hofmockel KS. 2014. Demonstrating microbial

- 1 co-occurrence pattern analyses within and between ecosystems.
- 2 Frontiers in microbiology 5:358-358.
- 3 53. Yilmaz B, Juillerat P, Oyas O, Ramon C, Bravo FD, Franc Y, Fournier N,
4 Michetti P, Mueller C, Geuking M, Pittet VEH, Maillard MH, Rogler G,
5 Wiest R, Stelling J, Macpherson AJ, Investigators SIC. 2019. Microbial
6 network disturbances in relapsing refractory Crohn's disease (vol 25, pg
7 323, 2019). Nature Medicine 25:701-701.
- 8 54. Cong J, Zhu J, Zhang C, Li T, Liu K, Liu D, Zhou N, Jiang M, Hou H,
9 Zhang X. 2019. Chemotherapy Alters the Phylogenetic Molecular
10 Ecological Networks of Intestinal Microbial Communities. Frontiers in
11 Microbiology 10:1008.
- 12 55. Kostic Aleksandar D, Chun E, Robertson L, Glickman Jonathan N,
13 Gallini Carey A, Michaud M, Clancy Thomas E, Chung Daniel C,
14 Lochhead P, Hold Georgina L, El-Omar Emad M, Brenner D, Fuchs
15 Charles S, Meyerson M, Garrett Wendy S. 2013. Fusobacterium
16 nucleatum Potentiates Intestinal Tumorigenesis and Modulates the
17 Tumor-Immune Microenvironment. Cell Host & Microbe 14:207-215.
- 18 56. Drewes JL, White JR, Dejea CM, Fathi P, lyadorai T, Vadivelu J,
19 Roslani AC, Wick EC, Mongodin EF, Loke MF, Thulasi K, Gan HM, Goh
20 KL, Chong HY, Kumar S, Wanyiri JW, Sears CL. 2017. High-resolution
21 bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal
22 common colorectal cancer consortia. npj Biofilms and Microbiomes

1 3:34.

2 57. Horiuchi A, Kokubu E, Warita T, Ishihara K. 2020. Synergistic biofilm
3 formation by *Parvimonas micra* and *Fusobacterium nucleatum*.
4 *Anaerobe* 62:102100.

5 58. Mark Welch JL, Ramírez-Puebla ST, Borisy GG. 2020. Oral Microbiome
6 Geography: Micron-Scale Habitat and Niche. *Cell Host & Microbe*
7 28:160-168.

8 59. Palmer RJ, Shah N, Valm A, Paster B, Dewhirst F, Inui T, Cisar JO.
9 2017. Interbacterial Adhesion Networks within Early Oral Biofilms of
10 Single Human Hosts. *Applied and Environmental Microbiology*
11 83:e00407-17.

12 60. Bowen WH, Burne RA, Wu H, Koo H. 2018. Oral Biofilms: Pathogens,
13 Matrix, and Polymicrobial Interactions in Microenvironments. *Trends in*
14 *Microbiology* 26:229-242.

15 61. Bengoechea JA, Sa Pessoa J. 2019. *Klebsiella pneumoniae* infection
16 biology: living to counteract host defences. *FEMS Microbiology*
17 *Reviews* 43:123-144.

18 62. Dinakaran V, Mandape SN, Shuba K, Pratap S, Sakhare SS, Tabatabai
19 MA, Smoot DT, Farmer-Dixon CM, Kesavalu LN, Adunyah SE,
20 Southerland JH, Gangula PR. 2019. Identification of Specific Oral and
21 Gut Pathogens in Full Thickness Colon of Colitis Patients: Implications
22 for Colon Motility. *Frontiers in Microbiology* 9:3220.

- 1 63. Sontag RL, Nakayasu ES, Brown RN, Niemann GS, Sydor MA,
2 Sanchez O, Ansong C, Lu S-Y, Choi H, Valleau D, Weitz KK,
3 Savchenko A, Cambronne ED, Adkins JN. 2016. Identification of Novel
4 Host Interactors of Effectors Secreted by Salmonella and Citrobacter.
5 mSystems 1:e00032-15.
- 6 64. Ananthakrishnan AN, Bernstein CN, Iliopoulos D, Macpherson A,
7 Neurath MF, Ali RAR, Vavricka SR, Fiocchi C. 2018. Environmental
8 triggers in IBD: a review of progress and evidence. Nat Rev
9 Gastroenterol Hepatol 15:39-49.
- 10 65. Dam B, Misra A, Banerjee S. 2019. Role of Gut Microbiota in
11 Combating Oxidative Stress, p 43-82. *In* Chakraborti S, Chakraborti T,
12 Chattopadhyay D, Shaha C (ed), Oxidative Stress in Microbial Diseases
13 doi:10.1007/978-981-13-8763-0_4. Springer Singapore, Singapore.
- 14 66. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC,
15 Funk K, Ketter A, Kim S, Kimchi A, Kitts PA, Kuznetsov A, Lathrop S, Lu
16 Z, McGarvey K, Madden TL, Murphy TD, O'Leary N, Phan L, Schneider
17 VA, Thibaud-Nissen F, Trawick BW, Pruitt KD, Ostell J. 2020. Database
18 resources of the National Center for Biotechnology Information. Nucleic
19 Acids Res 48:D9-D16.
- 20 67. Wu S, Sun C, Li Y, Wang T, Jia L, Lai S, Yang Y, Luo P, Dai D, Yang
21 YQ, Luo Q, Gao NL, Ning K, He LJ, Zhao XM, Chen WH. 2020. GMrepo:
22 a database of curated and consistently annotated human gut

- 1 metagenomes. *Nucleic Acids Res* 48:D545-D553.
- 2 68. Harrison PW, Ahamed A, Aslam R, Alako BTF, Burgin J, Buso N,
3 Courtot M, Fan J, Gupta D, Haseeb M, Holt S, Ibrahim T, Ivanov E,
4 Jayathilaka S, Balavenkataraman Kadhivelu V, Kumar M, Lopez R,
5 Kay S, Leinonen R, Liu X, O'Cathail C, Pakseresht A, Park Y, Pesant S,
6 Rahman N, Rajan J, Sokolov A, Vijayaraja S, Waheed Z, Zyoud A,
7 Burdett T, Cochrane G. 2021. The European Nucleotide Archive in
8 2020. *Nucleic Acids Res* 49:D82-D85.
- 9 69. Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R,
10 Goedert JJ, Shi J, Bork P, Sinha R. 2016. Colorectal Cancer and the
11 Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun
12 Sequencing. *PloS one* 11:e0155362-e0155362.
- 13 70. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G,
14 Almeida M, Arumugam M, Batto JM, Kennedy S, Leonard P, Li J,
15 Burgdorf K, Grarup N, Jorgensen T, Brandslund I, Nielsen HB, Juncker
16 AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J,
17 Tims S, Zoetendal EG, Brunak S, Clement K, Dore J, Kleerebezem M,
18 Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker JD,
19 Raes J, Hansen T, Meta HITc, Bork P, Wang J, Ehrlich SD, Pedersen O.
20 2013. Richness of human gut microbiome correlates with metabolic
21 markers. *Nature* 500:541-6.
- 22 71. Schirmer M, Franzosa EA, Lloyd-Price J, McIver LJ, Schwager R, Poon

- 1 TW, Ananthakrishnan AN, Andrews E, Barron G, Lake K, Prasad M,
- 2 Sauk J, Stevens B, Wilson RG, Braun J, Denson LA, Kugathasan S,
- 3 McGovern DPB, Vlamakis H, Xavier RJ, Huttenhower C. 2018.
- 4 Dynamics of metatranscription in the inflammatory bowel disease gut
- 5 microbiome. *Nat Microbiol* 3:337-346.
- 6 72. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer
- 7 for Illumina sequence data. *Bioinformatics* 30:2114-20.
- 8 73. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with
- 9 Bowtie 2. *Nat Methods* 9:357-9.
- 10 74. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O,
- 11 Huttenhower C. 2012. Metagenomic microbial community profiling
- 12 using unique clade-specific marker genes. *Nature Methods* 9:811-+.
- 13 75. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M,
- 14 Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N,
- 15 Huttenhower C. 2018. Species-level functional profiling of
- 16 metagenomes and metatranscriptomes. *Nature Methods* 15:962-+.
- 17 76. Subramanian B, Gao S, Lercher MJ, Hu S, Chen WH. 2019. Evolview
- 18 v3: a webserver for visualization, annotation, and management of
- 19 phylogenetic trees. *Nucleic Acids Res* 47:W270-W275.
- 20 77. Friedman J, Alm EJ. 2012. Inferring Correlation Networks from
- 21 Genomic Survey Data. *PLOS Computational Biology* 8:e1002687.
- 22 78. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2010. A basic

1 introduction to fixed-effect and random-effects models for meta-analysis.
 2 Research Synthesis Methods 1:97-111.
 3 79. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. 2019. Cytoscape
 4 Automation: empowering workflow-based network analysis. Genome
 5 Biol 20:185.

7 FIGURES LEGENDS

8 Fig. 1. Disease-specific and shared microbial markers showed distinct
 9 prevalence profiles in patients and controls. A. Microbial markers and their
 10 trends (i.e. case- or control- enriched) in common intestinal diseases. Species
 11 significantly enriched in cases (or controls) of corresponding disease in
 12 meta-analysis are shown ($\text{fdr} < 0.05$, Benjamini-Hochberg FDR correction),
 13 with their phylum shown on top. Red indicates case-enriched species, and
 14 blue indicates control-enriched one. B. The boxplot shows the inner Jaccard
 15 similarities of case-enriched microbes in all cases. The case-enriched
 16 microbes were clustered according to their trends in intestinal diseases (see
 17 Materials and Methods). The background means the similarities between
 18 members that did not belong to the same cluster. The four clusters were
 19 named according to their members; UC-specific, CD-specific and
 20 CRC-specific clusters only contained the disease-specific markers, while
 21 common-enriched cluster contained markers from at least two diseases (****: p
 22 < 0.0001). C. The barplot shows the fraction of cases that are “positive” for

1 given clusters in per disease type. Here positive samples for a given species
2 are defined as those in which the species was found with higher relative
3 abundance than 95% of all controls. The significant differences of the positive
4 fraction between controls and cases for each cluster were assessed via
5 Crochran-Mantel-Haenszel test with ‘dataset’ as the stratified factor; the ‘star’
6 symbols on the bars were used to indicate that marker species of a given
7 cluster were significantly more prevalent in cases than in their corresponding
8 controls (*: $p < 0.05$, **: $p < 0.01$).

9 **Fig. 2. Disease-specific pattern of metagenomics functional modules. A.**

10 Functional modules were summed according to the category of MetaCyc
11 database. The differences between controls and cases in one specified
12 disease were calculated as the generalized fold changes, and the
13 significances were assessed using two-sided Wilcoxon rank sum test and
14 blocked with “datasets” (see Materials and Methods). Red indicates
15 case-enriched modules, and blue indicates control-enriched modules. The
16 asterisk indicated the module were significantly different between cases and
17 controls (*: $p < 0.05$, **: $p < 0.01$). B. The boxplots show the distribution of
18 some representative modules in the three intestinal diseases. The asterisks on
19 the bars were calculated as mentioned above, indicating the significances.

20 **Fig. 3. Marker microbes signified distinct degradation preferences. Shown**

21 here are the correlations of meta-analysis in relative abundances between the
22 degradation pathways and the marker species (see Materials and Methods).

1 The pathways were clustered using 'mcquitty' algorithm, while the species
2 were sorted by their related diseases and changing trends. The blocks in the
3 heatmap show the overall coefficients from the meta-analysis. The red block
4 indicates positive correlation and blue indicates negative one. The asterisk
5 indicates adjusted P-value of the overall coefficients in the meta-analysis is
6 below 0.05. The similar plot for other pathways is shown in Fig. S5.

7 **Fig. 4. Increased correlations of marker species in diseased conditions. A-B.**

8 The statistics on the increases of positive correlated pairs (A) and negative
9 correlated pairs (B) in case network (red) as compared with control network
10 (blue) in each of the three diseases. Y-axis shows the case-to-control ratios of
11 each disease, with number of pairs in the control network being normalized to
12 1. The numbers in the bars indicate the actual numbers of correlated pairs.

13 **C-E. Detected modules from correlation networks among marker species of**
14 **per state in corresponding disease datasets (see Materials and Methods).**
15 Color of nodes means the alteration trends of species, and the sizes mean the
16 prevalence of the bacteria in the overall health (or cases) of given datasets.
17 Gray edges indicate positive relationship and red edges indicate negative.
18 Thickness of edges indicates correlation strength.

19 **Fig. 5. Random forest models for patient stratification using taxonomic and/or**
20 **metabolic profiles. A.** Binary models for distinguishing cases from controls
21 were constructed using various type of features as shown in plot. The
22 classification results were plotted as the AUROCs of corresponding model.

1 AUROC is evaluated through ten-time ten-fold random forest cross-validation
2 in all samples (see Materials and Methods). The combined profile indicated the
3 relative abundance profile of all taxonomic and metabolic features. The
4 combined markers indicated the relative abundance of all taxonomic and
5 metabolic markers. B-D. Confusion matrix evaluation of the four-class model
6 (B), three-class model (C), and cases model (D) based on combined marker
7 features for distinguishing different physical condition. The number in row *i* and
8 column *j* in the matrix on the left part indicates how many patients with disease
9 *i* actually were categorized to disease *j* in model. Color filling the cell means
10 the relative size of the number in the corresponding row. The right part is the
11 TPR for per disease type. Total accuracy indicates the fraction of all correct
12 predictions, and 95% CI is the confidence interval of accuracy.

13

14 Supplemental Material

15 Table S1. A list of CRC and IBD datasets used in the study and statistical
16 results of confounding factors, HDCs and Shannon diversity. The project
17 "SRP057027" lacks detailed information about age, BMI and gender, but the
18 original literature gave a statistical table which shows that there were no
19 significant differences in age and gender between controls and CD patients
20 (24). Numerical variables were reported as the median values in the
21 corresponding group. The P-values were calculated using Wilcoxon rank sum
22 test or chi-squared test to compare cases to controls.

1 **Fig. S1. The pipeline for metagenomics data collection.** We collected records
2 about IBD and CRC in public databases until Sep, 2019. The requirements for
3 samples were that (1) the number of each group in each dataset was not
4 below 20, and (2) there were no overlapping samples across datasets.

5 **Fig. S2. The associated effects of identified marker species in each dataset of**
6 **each disease.** The marker species which were also significantly differential
7 species in corresponding dataset ($P\text{-value} < 0.05$) were plot. Each block
8 indicates a marker species of corresponding disease. Red means that the
9 case-enriched species and blue means the control-enriched one. A, CRC; B,
10 CD; C, UC.

11 **Fig. S3. Phylogenetic relationships of disease-specific and shared marker**
12 **species.** Phylogenetic relationships of the marker species were based on the
13 NCBI common tree, generated using the online phyloT tool
14 (<https://phylot.biobyte.de>), and visualized using Evolview ver3.0 (76) (with
15 manual annotations). The LCAs were determined according to the NCBI
16 taxonomy information of the species in corresponding branch.

17 **Fig. S4. Identified marker pathways that showed consistent changes in the**
18 **respective diseases.** The relative abundances of pathways were identified
19 using HuManN2 on the metagenomics data. Pathways significantly enriched in
20 cases (or controls) of corresponding disease in meta-analysis are shown ($\text{fdr} <$
21 0.05 , Benjamini-Hochberg FDR correction, see Materials and Methods). Red
22 block indicates case-enriched pathway, and blue block indicates

1 control-enriched one.

2 **Fig. S5. Marker microbes showed distinct biosynthesis and other metabolic**
 3 **pathways preferences.** Similar to Fig. 3, however, shown here are the
 4 correlations of meta-analysis in relative abundances between the
 5 non-degradation pathways and the marker species (see Materials and
 6 Methods). The pathways were clustered using 'mcquitty' algorithm, while the
 7 species were sorted by their related diseases and changing trends. The blocks
 8 in the heatmap show the overall coefficients from the meta-analysis. The red
 9 block indicates positive correlation and blue indicates negative one. The
 10 asterisk indicates adjusted P-value of the coefficient in meta-analysis is below
 11 0.05.

12 **Fig. S6. Some marker species and pathways correlated significantly with**
 13 **HDCs.** Each block indicates the marker species (A) and marker species (B)
 14 identified in meta-analysis. Red means that the species/pathways was
 15 HDC-related evidently in at least two datasets of corresponding disease type.
 16 Blue means the species/pathways was not HDC-related.

17 **Fig. S7. Correlation networks among disease-altered species in CRC datasets**
 18 **(A-B), CD datasets (C-D) and UC datasets (E-F).** Color of nodes means the
 19 alteration trends of species in its corresponding datasets. The sizes mean the
 20 prevalence of the bacteria in the overall controls (or cases) of corresponding
 21 datasets. Gray edges indicate positive relationship and red edges indicate
 22 negative. Thickness of edges indicates correlation strength.

1 Fig. S8. Evaluation for ECSs and BCSs of networks showed in Fig. S7. Blue
2 bars indicate the centralities of control network from datasets, and red bars
3 indicate centralities in case network. A-B, CRC; C-D, CD; E-F, UC.

4 Fig. S9. Cross-validation models (A-B) and LODO models (C) based on
5 various types of features for different purposes. A: The heatmap depicts the
6 cross-validation results of multi-class models built on corresponding features.
7 Rownames indicate the classification tasks (Four-class: CTR/UC/CD/CRC;
8 Three-class: CTR/IBD/CRC; Cases: UC/CD/CRC). Column names indicate the
9 type of features. Each block indicates the accuracy of each model. B:
10 Confusion matrix evaluation of four-class model built on combined profile in
11 taxonomic and functional level for distinguishing different disease and controls.
12 The number in row i and column j in the matrix on the left part indicates how
13 many samples with state i actually were categorized to state j in model. Color
14 filling the cell means the relative size of the number in the corresponding row.
15 The right part is the TPR for per physical condition. Total accuracy indicates
16 the fraction of all correct predictions, and 95% CI is the confidence interval of
17 accuracy. C: The heatmap depicts the training results and validation result
18 of each LODO model. Each row indicates the result of the LODO models
19 training in the remainig data except the appointed dataset (see Materials and
20 Methods). Column names were the combination of the classification task and
21 type of features. For example, the 'Four-class All' means the LODO models
22 based on combined profile were used to distinguish four classes

1 (CTR/UC/CD/CRC), when the 'Four-class Dif' means means the LODO
2 models based on combined markers to distinguish four classes. In this plot we
3 only showed the models based on either combined profile (All) or combined
4 markers (Dif). For the multi-class models, including the models starting with
5 'Four-class', 'Three-class' and 'Cases', we used the accuracy as the
6 evaluation of classification. For the binary models, including 'Binary All' and
7 'Binary Dif', were displayed as AUROCs. The model average were the mean
8 values of each column. The LODO validation were the integrated validation
9 results of the LODO models in the same column.

10

A **B**

A

B

C

D