

Supplementary Information

Faecal microbiome-based machine learning for multi-class disease diagnosis

Qi Su, Qin Liu, Raphaela Iris Lau, Jingwan Zhang, Zhilu Xu, Yun Kit Yeoh, Thomas WH Leung, Whitney Tang, Lin Zhang, Jessie QY Liang, Yuk Kam Yau, Jiaying Zheng, Chengyu Liu, Mengjing Zhang, Chun Pan Cheung, Jessica YL Ching, Hein M Tun, Jun Yu, Francis KL Chan, Siew C Ng*

*Correspondence:

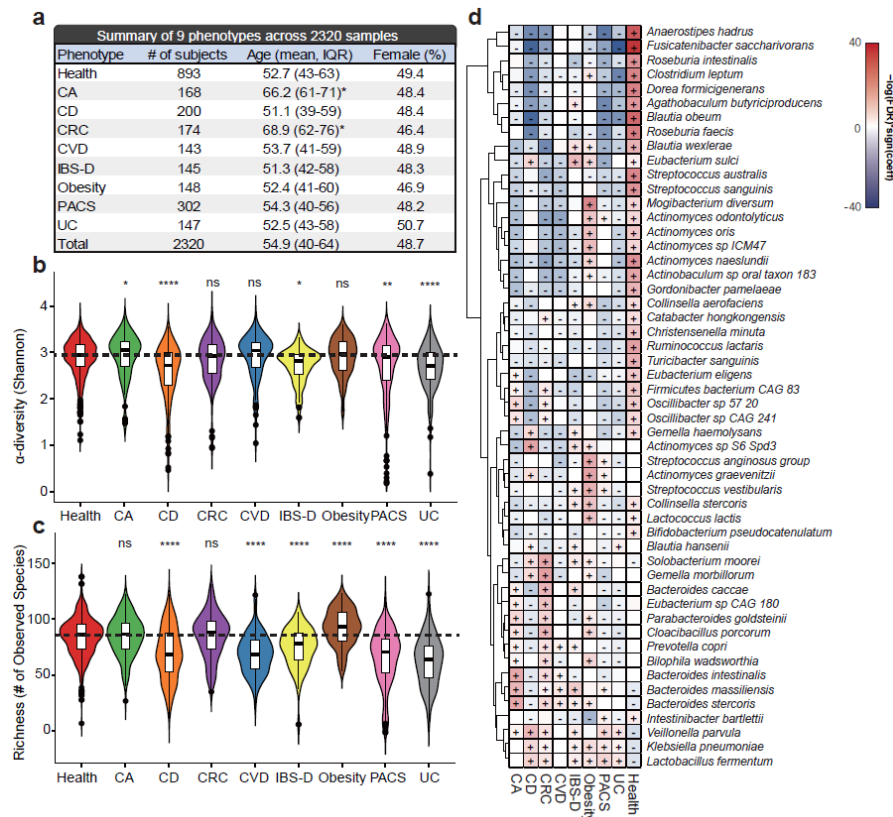
Siew Chien Ng, Department of Medicine and Therapeutics, Faculty of Medicine, The Chinese University of Hong Kong.

E-mail: siewchienng@cuhk.edu.hk

Tel: (852)3505-1506

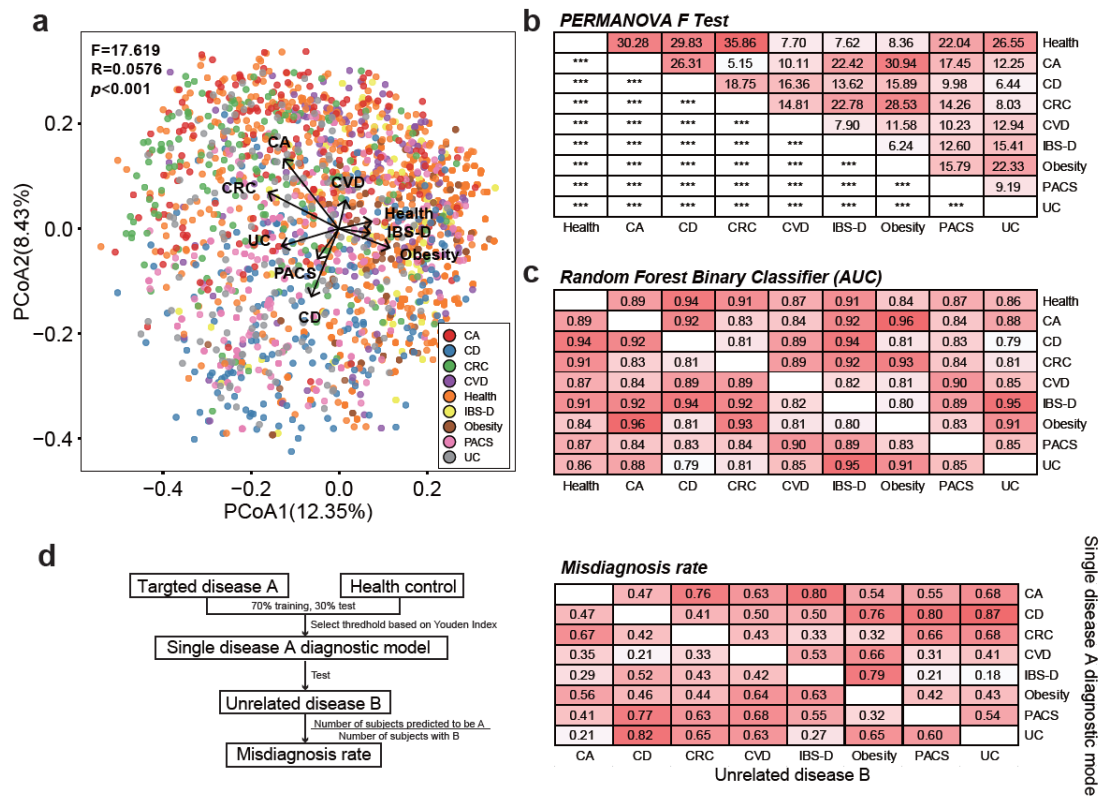
Fax: (852) 2647-3852

Supplementary Figures

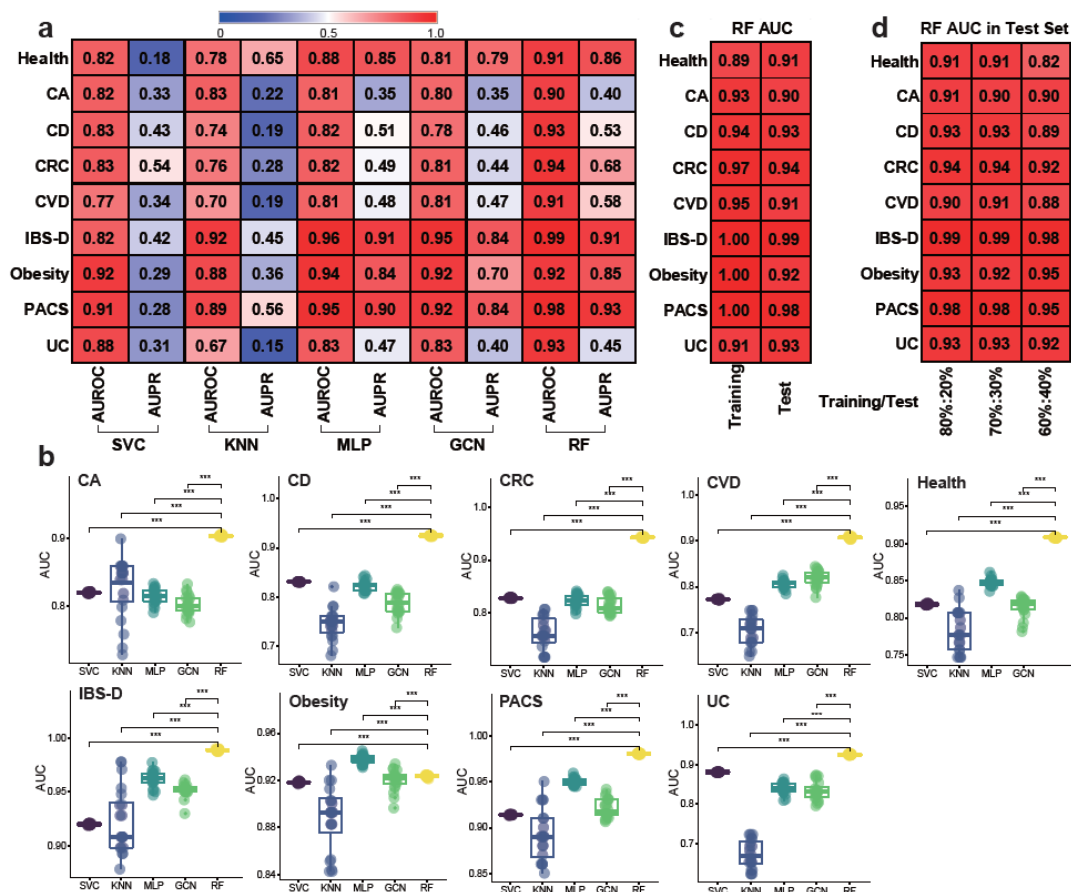


Supplementary Figure 1 Microbiome signatures of health and diseases.

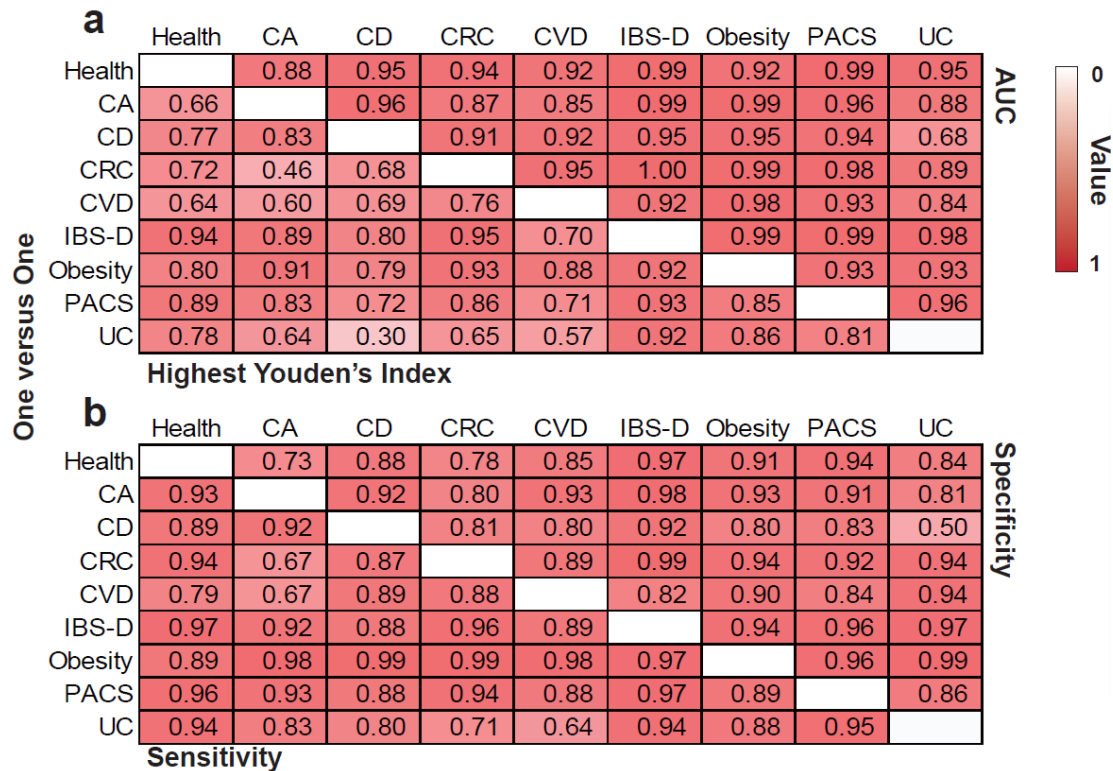
a, Demographics of subjects recruited in this study. Subjects with CA or CRC are older than subjects in other phenotypes ($p < 0.05$, Kruskal–Wallis test), but no difference between CA and CRC. No difference was found in gender among different phenotypes. **b**, Alpha diversity (Shannon index) and **c**, Richness (number of observed microbial species) in different phenotypes. P values were calculated by comparing each disease and healthy subjects using MaAsLin2 after adjustment of age, gender and sample processing. Health ($n=893$), CA ($n=168$), CD ($n=200$), CRC ($n=174$), CVD ($n=143$), IBS-D ($n=145$), PACS ($n=302$), UC ($n=147$). Centre line is the median, box limits indicate upper and lower quartiles, whiskers show $1.5 \times$ interquartile range, points indicate outliers and the outline displays the distribution of the data. **d**, Heat map of microbial species associated with health and mixed diseases. The top 50 microbial species with the highest number of associations were visualised. Associations were coloured by direction of effect (red, positive; blue, negative; $p < 0.05$), with associations significant at $FDR < 0.05$ marked with a plus (positive correlations) or minus (negative correlations), respectively. The nominal significance (p value) of associations were calculated by MaAsLin 2, and the FDR was computed by Benjamini-Hochberg correction. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis. Source data are provided as a Source Data file.



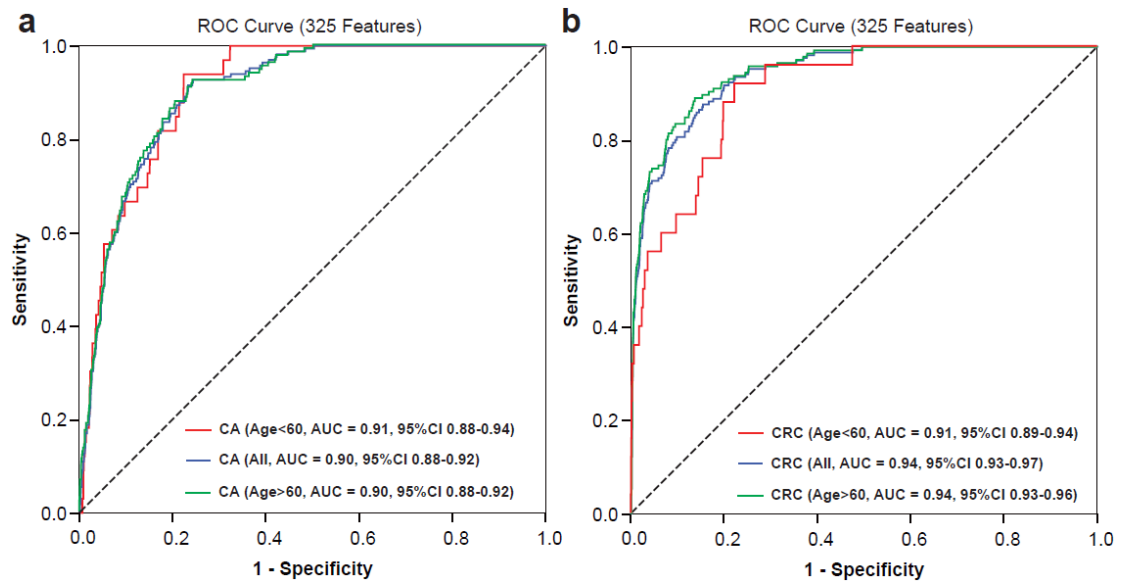
Supplementary Figure 2 Faecal microbiome differences among different phenotypes. **a**, Bray-Curtis dissimilarity-based PCoA analysis of species-level significantly separated different phenotypes. F, R and P values were calculated by PERMANOVA test with 999 permutations. **b**, Beta-diversity-based F statistics for one phenotype versus one another. F and P values were calculated by PERMANOVA test with 999 permutations. **c**, Area under the receiver operating characteristic curve (AUC) of random forest binary classifiers for one versus one discrimination of multiple phenotypes. **d**, misdiagnosis rate of single disease models in unrelated disease phenotypes. Source data are provided as a Source Data file. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis.



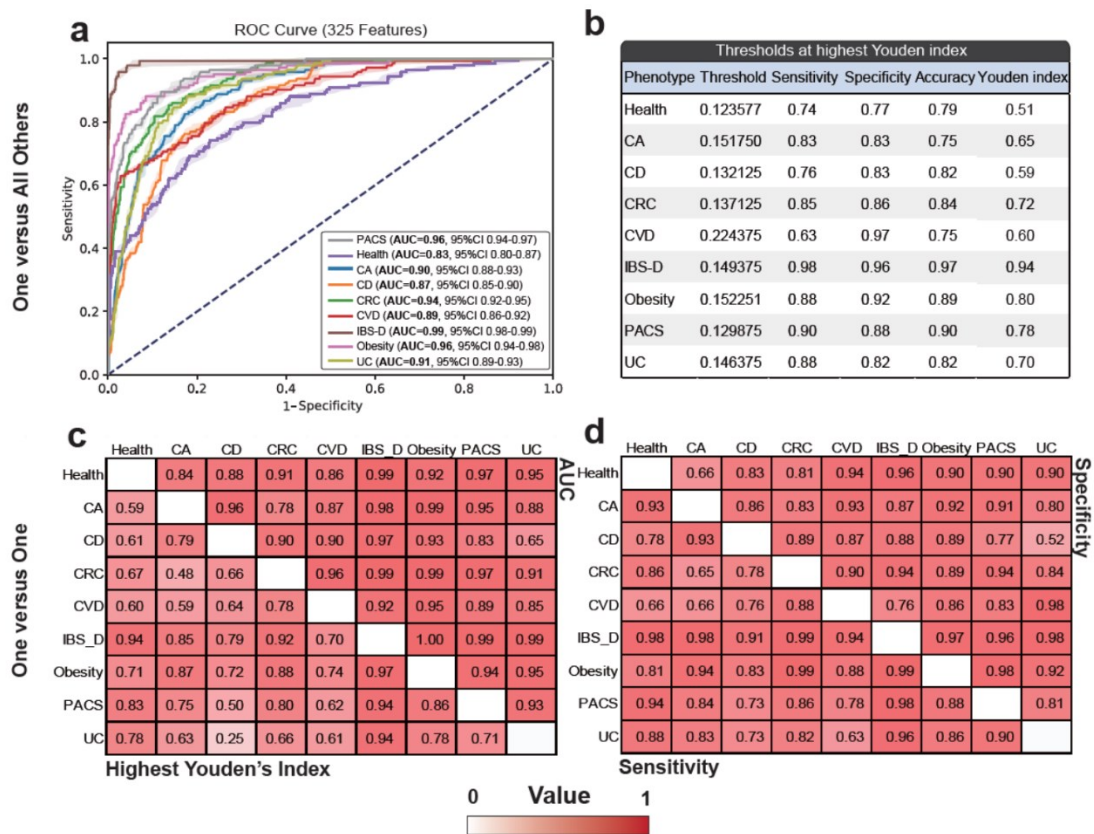
Supplementary Figure 3 Comparison of different classifiers and split ratio for multi-class phenotypes classification using faecal microbiome data at the species level. **a**, The performance across models was measured using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) for predicting one phenotype versus all others in the independent test set. **b**, Comparison of AUROC (n=20 repeats) for predicting one phenotype versus all others by different models. P values were calculated by Kruskal–Wallis test. *** represents $p < 0.001$. **c**, AUROC of RF multi-class model in the training set (calculated by 5-fold cross-validation) and test set. Centre line is the median, box limits indicate upper and lower quartiles, whiskers show $1.5 \times$ interquartile range, and points indicate outliers. **d**, Performance of the Random Forest multi-class model across different split ratios. SVM, support vector machine; KNN, K-nearest neighbours; RF, Random forests; MLP, multi-layer perceptron; GCN, graph convolutional neural network; CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis. Source data are provided as a Source Data file.



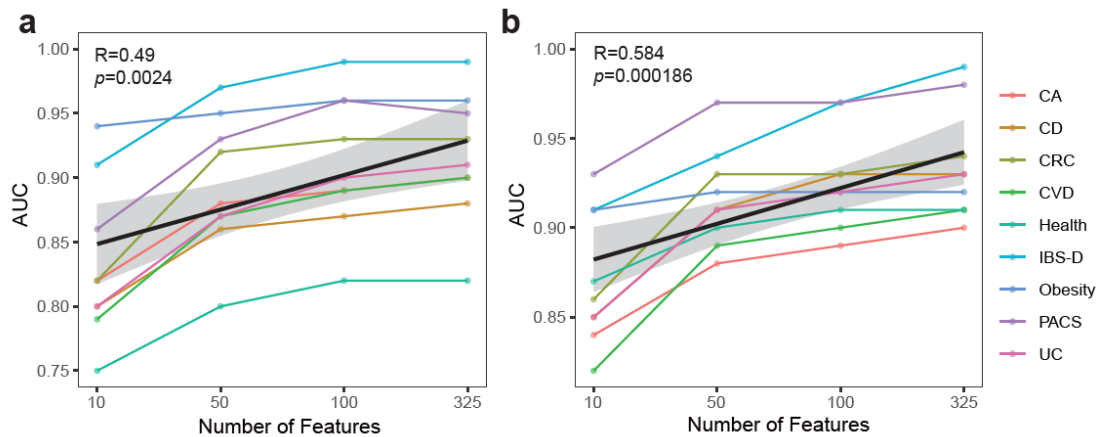
Supplementary Figure 4 Performance of random forest multi-class classifier in one versus one discrimination of multiple phenotypes. a, Area under the receiver operating characteristic curve (AUC) and the highest Youden's index of the random forest multi-class classifier using species-level faecal microbiome data. **b**, sensitivities and specificities selected based on the highest Youden's index for one versus one discrimination of multiple phenotypes. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis. Source data are provided as a Source Data file.



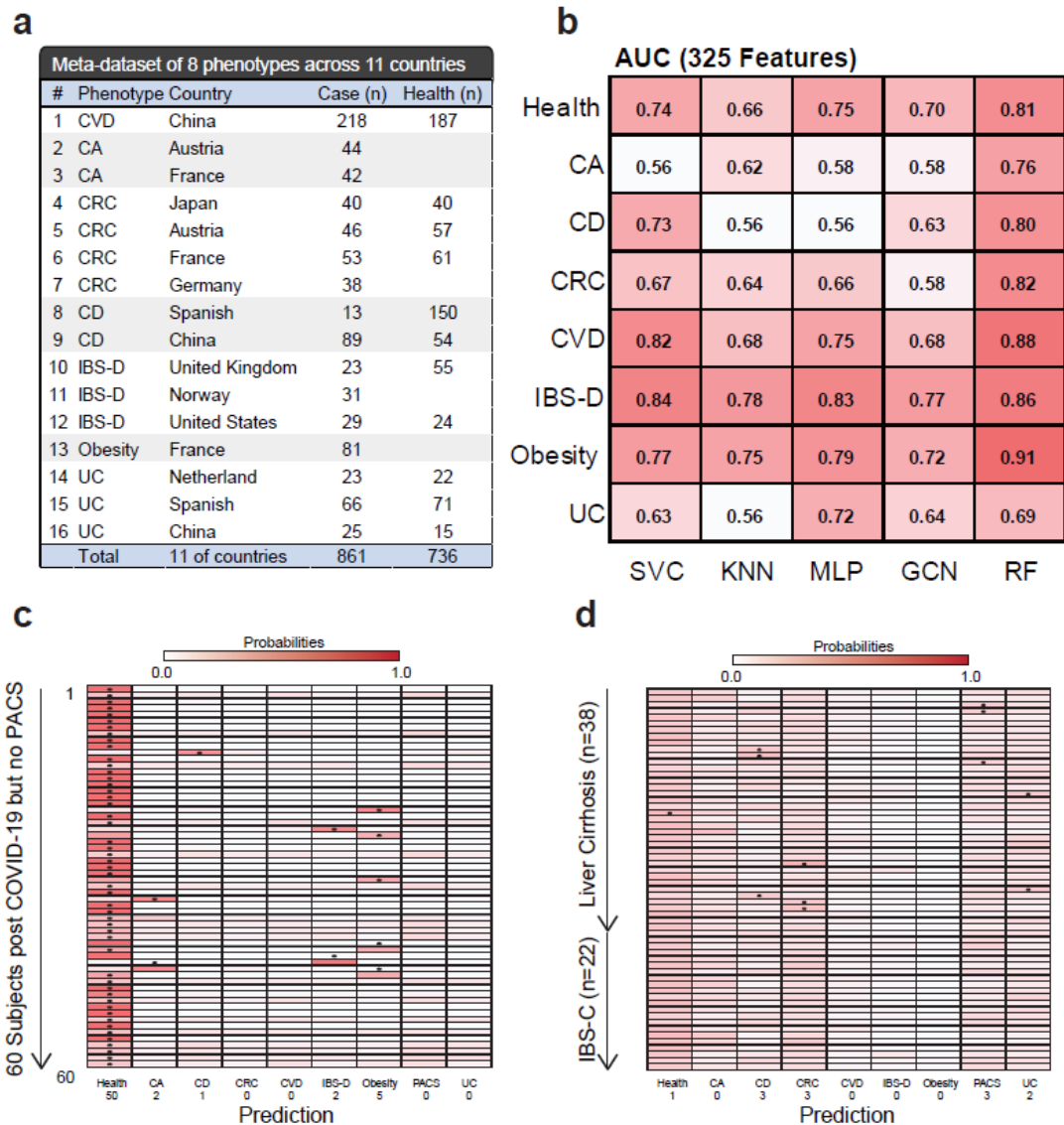
Supplementary Figure 5 Performance of random forest multi-class classifier for colorectal adenomas (a) and colorectal cancer (b) samples stratified by age. CA, colorectal adenomas; CRC, colorectal cancer. Source data are provided as a Source Data file.



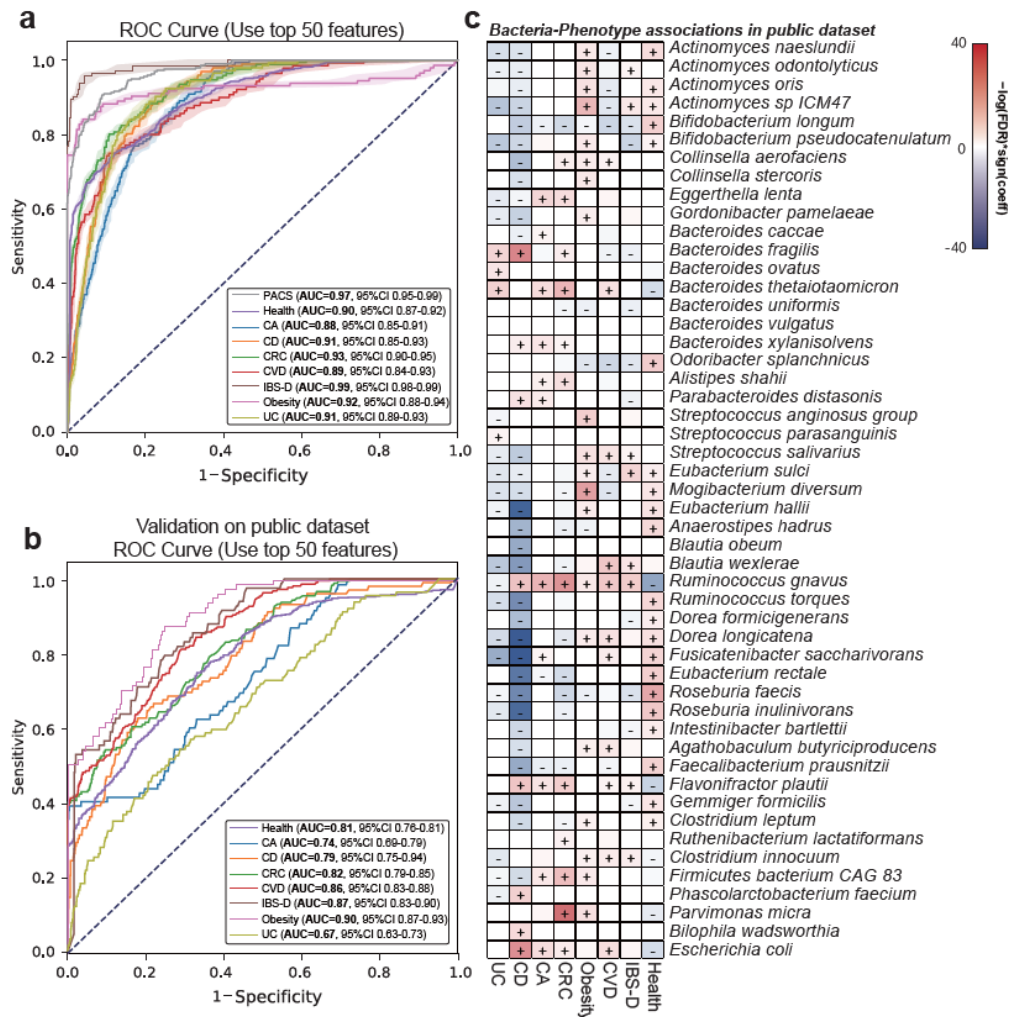
Supplementary Figure 6 Performance of random forest multi-class classifier trained and tested in a balanced sample size cohort. **a**, Area under the receiver operating characteristic curve (AUROC, centre for the error bands is median) and **b**, Model performance metrics details of random forest multi-class classifier for diagnosing one phenotype versus all others using species-level faecal microbiome data. **c**, AUROC and the highest Youden's index of the random forest multi-class classifier for one versus one discrimination of multiple phenotypes. **d**, Sensitivities and specificities selected based on the highest Youden's index for one versus one discrimination of multiple phenotypes. 143 subjects for each phenotype were randomly selected to form a balanced sample size cohort (a total of 1,287, 70% for training and 30% for test). CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis. Source data are provided as a Source Data file.



Supplementary Figure 7 Performance of random forest multi-class classifier using a different number of features. The area under the receiver operating characteristic curve (AUROC) of random forest multi-class classifier for diagnosing one phenotype versus all others using the top 10, 50, 100 and all 325 features in (a) the balanced sample size cohort (n=1,287) and (b) the complete cohort (n=2,320). Error bands reflect the 95%CI. R and P values were calculated by Spearman Correlation. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis. Source data are provided as a Source Data file.



Supplementary Figure 8 Independent validations of the performance of random forest multi-class classifier. **a**, Summary of involved phenotypes and corresponding sample size and source countries in the publicly available cross-regional datasets. **b**, Area under the receiver operating characteristic curve (AUROC) of the trained random forest multi-class classifier for diagnosing one phenotype versus all others in the cross-regional dataset. **c**, Probabilities yielded by the trained random forest multi-class classifier for 60 subjects who completely recovered from COVID-19 without any post-acute COVID-19 syndrome. **d**, Probabilities yielded by the trained random forest multi-class classifier for 60 subjects with liver cirrhosis (n=38) or constipation-dominant IBS (n=22). The highest probability for each subject that exceed corresponding thresholds are marked with an asterisk. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis. Source data are provided as a Source Data file.



Supplementary Figure 9 Performance of random forest multi-class classifier using top 50 features. **a**, Area under the receiver operating characteristic curve (AUROC, centre for the error bands is median) of random forest multi-class classifier (using top 50 features) for diagnosing one phenotype versus all others in the independent test set. **b**, Validation of the performance of random forest multi-class classifier (using top 50 features) in the publicly available datasets. **c**, Microbial species associated with health status or different disease phenotypes in the assembled public dataset. The top 50 microbial species contributing to the random forest multi-class classifier were visualized. Associations were coloured by direction of effect (red, positive; blue, negative; $p < 0.05$), with associations significant at $\text{FDR} < 0.05$ marked with a plus (positive correlations) or minus (negative correlations), respectively. The nominal significance (p value) of associations were calculated by MaAsLin 2, and the FDR was computed by Benjamini-Hochberg correction. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; UC, ulcerative colitis. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis. Source data are provided as a Source Data file.

Supplementary Tables

Supplementary Table 1 Beta-diversity-based F statistics (PERMANOVA test) for one phenotype versus another. Part of older subjects with CA and CRC was excluded to eliminate the age difference. F and P values were calculated by PERMANOVA test with 999 permutations.

	8.29	29.83	10.26	7.70	7.62	8.36	22.04	26.55	Health
***		9.84	5.15	4.71	9.37	12.79	6.27	5.43	CA
***	***		5.42	16.36	13.62	15.89	9.98	6.44	CD
***	***	***		5.45	7.92	8.85	5.52	2.36	CRC
***	***	***	***		7.90	11.58	10.23	12.94	CVD
***	***	***	***	***		6.24	12.60	15.41	IBS-D
***	***	***	***	***	***		15.79	22.33	Obesity
***	***	***	***	***	***	***		9.19	PACS
***	***	***	***	***	***	***	***		UC

Notes: *** means $p < 0.001$. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; UC, ulcerative colitis. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis.

Supplementary Table 2 Summary of the publicly available metagenome datasets

#	Phenotypes	Country	Year	Author	Case (n)	Control (n)	Data Accession Number
1	UC	Netherland	2018	Franzosa EA et al.	23	22	PRJNA400072
2	UC	Danish/Spanish	2014	Nielsen HB et al.	66	221	PRJEB1220
3	UC	China	2019	Weng YJ et al.	25	15	PRJNA429990
4	CD	Danish/Spanish	2014	Nielsen HB et al.	13		PRJEB1220
5	CD	China	2019	Weng YJ et al.	41		PRJNA429990
6	CD	China	2017	He Q et al.	48	54	PRJEB15371
7	CVD	China	2017	Jie et al.	218	187	ERP023788
8	CRC	Japan	2019	Yachlda et al.	40	40	DRA006684, DRA008156
9	CRC	Austria	2015	Feng et al.	46	57	ERP008729
10	CRC	France/Germany	2014	Zeller et al.	91	61	ERP005534
11	IBS-D	United Kingdom	2021	Vervier K et al.	23	55	ERP021923
12	IBS-D	Norway	2020	Goll et al.	31		PRJEB36140
13	IBS-D	United States	2020	Mars et al.	29	24	PRJEB37924
14	Obesity	France	2020	Meslier et al.	81		PRJEB33500
15	Adenoma	Austria	2015	Feng et al.	44		ERP008729
16	Adenoma	France	2014	Zeller et al.	42		ERP005534

Notes: CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; UC, ulcerative colitis. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD,

Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis.

Supplementary Table 3 Performance details of the trained random forest multi-class classifier using publicly available datasets

Phenotype	Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
CA	0.758	0.025	7.690E-16	0.709	0.807
CD	0.798	0.022	5.845E-24	0.755	0.842
CRC	0.819	0.016	1.286E-43	0.787	0.851
CVD	0.879	0.012	1.958E-72	0.856	0.902
Health	0.811	0.011	5.578E-102	0.790	0.832
IBS_D	0.858	0.019	3.653E-28	0.821	0.895
Obesity	0.913	0.013	7.359E-37	0.893	0.942
UC	0.693	0.026	6.806E-12	0.641	0.744

Notes: CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; UC, ulcerative colitis. CA, colorectal adenomas; CD, Crohn's disease; CRC, colorectal cancer; CVD, Cardiovascular disease; IBS-D, diarrhea-dominant irritable bowel syndrome; PACS, post-acute COVID-19 syndrome; UC, ulcerative colitis.

Supplementary Table 4 Importance of scores of the top 50 microbial species in the trained random forest multi-class classifier

Feature	Importance
<i>Blautia_wexlerae</i>	0.01307182
<i>Fusicatenibacter_saccharivorans</i>	0.012884895
<i>Bacteroides_vulgatus</i>	0.011013169
<i>Agathobaculum_butyriciproducens</i>	0.01085556
<i>Dorea_longicatena</i>	0.010582717
<i>Eubacterium_hallii</i>	0.01009724
<i>Faecalibacterium_prausnitzii</i>	0.009771901
<i>Bacteroides_thetaiotaomicron</i>	0.009741041
<i>Bacteroides_uniformis</i>	0.009703679
<i>Escherichia_coli</i>	0.009688801
<i>Collinsella_aerofaciens</i>	0.009548232
<i>Streptococcus_salivarius</i>	0.009435219
<i>Parvimonas_micra</i>	0.009415211
<i>Eubacterium_sulci</i>	0.009327508
<i>Blautia_obeum</i>	0.009163572

<i>Actinomyces_odontolyticus</i>	0.008993389
<i>Parabacteroides_distasonis</i>	0.008778252
<i>Anaerostipes_hadrus</i>	0.008766194
<i>Flavonifractor_plautii</i>	0.008702062
<i>Bacteroides_fragilis</i>	0.008668667
<i>Mogibacterium_diversum</i>	0.008467032
<i>Dorea_formicigenerans</i>	0.008148955
<i>Ruthenibacterium_lactatiformans</i>	0.007992084
<i>Bilophila_wadsworthia</i>	0.007902528
<i>Roseburia_inulinivorans</i>	0.007901513
<i>Ruminococcus_gnavus</i>	0.007874336
<i>Actinomyces_naeslundii</i>	0.007748365
<i>Bifidobacterium_longum</i>	0.007733054
<i>Actinomyces_sp_ICM47</i>	0.007620109
<i>Gordonibacter_pamelaeae</i>	0.007555373
<i>Intestinibacter_bartlettii</i>	0.007491595
<i>Eubacterium_rectale</i>	0.007419018
<i>Collinsella_stercoris</i>	0.007119509
<i>Eggerthella_lenta</i>	0.007105053
<i>Roseburia_faecis</i>	0.007067439
<i>Streptococcus_parasanguinis</i>	0.007066339
<i>Ruminococcus_torques</i>	0.007000123
<i>Phascolarctobacterium_faecium</i>	0.006936069
<i>Bacteroides_ovatus</i>	0.006908834
<i>Bacteroides_xylanisolvens</i>	0.006900542
<i>Bifidobacterium_pseudocatenulatum</i>	0.006819506
<i>Clostridium_leptum</i>	0.006804848
<i>Actinomyces_oris</i>	0.006801588
<i>Firmicutes_bacterium_CAG_83</i>	0.006658731
<i>Streptococcus_anginosus_group</i>	0.006546908
<i>Odoribacter_splanchnicus</i>	0.006539549
<i>Gemmiger_formicilis</i>	0.006414319
<i>Alistipes_shahii</i>	0.006407032
<i>Bacteroides_caccae</i>	0.006356583

Supplementary Table 5 Questionnaire used for post-acute COVID-19 symptom assessment

Symptoms	Yes/No	Duration
Fever		
Chills		
Cough		
Sputum Production		
Sore throat		
Congested or runny nose		
Fatigue		
Joint pain		
Muscle pain		
Shortness of breath		
Headache		
Dizziness		
Nausea		
Vomiting		
Diarrhoea		
Loss of taste		
Loss of smell		
Abdominal pain		
Epigastric pain		
Difficulty in concentration		
Inability to exercise		
Difficulty in sleeping		
Anxiety		
Sadness		
Memory problem		
Chest pain		
Palpitations		
Night sweats		
Hair loss		
Blurred vision		
Any other symptoms		