nature portfolio

Corresponding author(s): Siew Chien Ng and Jingwan Zhang

Last updated by author(s): Aug 12, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For full details see Methods. Metagenomic sequencing data was carried out on the Illumina NovaSeq 6000 platform. Multiplex droplet digital PCR data were generated on the Bio-rad QX200 system. Other sequencing data were downloaded from the public datasets.
Data analysis	CuratedMetagenomicData (v3.10.0), Trimmomatic(v0.39), Kneaddata (v0.10.0), MetaPhlAn3 (v3.0.13), Bowtie2 (v2.4.2), HUMAnN (v3.0) were used to acquire and process microbiome sequencing data. Other software used for data analysis and visualization: R version 4.0.2, R package: vegan (v2.6.4), MaAsLin2 (v1.4.0), RandomForest (v4.6-14), caret (v6.0-94), pROC (v1.18.2), kernelshap (v0.3-7), shapviz (v0.8-0), phyloseq(v1.34.0), pheatmap(1.0.12), ggpubr(v0.6.0), ggplot2(v3.4.4).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The metagenomics data generated in this study have been deposited in the National Center for Bioinformation database under accession code PRJNA1086048 (https://www.ncbi.nlm.nih.gov/sra/PRJNA1086048). Other raw metagenomic data are available in the SRA and European Nucleotide Archive under accession number PRJNA400072, PRJNA429990, PRJEB15371, and PRJEB37924. Metagenomic taxonomy profiling data from Hall AB 2017 cohort (PRJNA385949), Nielsen HB 2014 cohort (PRJEB1220), HMP 2019 IBDMDB cohort (PRJNA398089), LifeLD VilaAV 2018 cohort (EGAS00001001704, EGAD00001004194), and Ijaz UZ 2017 (PRJEB18780), and CRC and CA cohorts (Feng Q, Zeller G, Wirbel J, Yachida S), and obesity cohort from LeChatelierE 2013) was acquired from curatedMetagenomicData (v3.10.0). Human reference genome (hg37decv0.1) were downloaded from https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.25_GRCh37.p13/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	In this study, we did not have any specific requirements for participants' gender. The information of gender was collected based on self-reporting. All subjects have provided written informed consent. The gender of participants in discover cohort, validation cohorts and public datasets were described separately in the Supplementary Table S1 and Supplementary Table S2.
Reporting on race, ethnicity, or other socially relevant groupings	The discovery cohort in this study only used the Hong Kong population, and the validation cohorts used the Hong Kong and Australian populations. Public datasets from different regions (including the United States, Netherlands, Mainland China, Spain, Denmark, and the United Kingdom) were used to verify the stability of the diagnostic model. Non-IBD patients from Hong Kong, Austria, France, Germany, Japan, the United States, and Denmark were also used to validate the accuracy of the diagnostic model. Independent cohorts from Canada and Taiwan were used for validation using the multiplex droplet digital PCR method. The race, ethnicity, or other socially relevant groupings of participants in this study was described in the Supplementary Table S2.
Population characteristics	A total of 5,979 samples, including 1,884 samples from in-house sequencing datasets and 4,095 samples from public datasets, was included in this study. In discovery cohort, a total of 497 subjects from Hong Kong were enrolled, comprising 174 CD patients, 205 UC patients and 118 controls. The Hong Kong validation cohort includes 139 UC patients and 139 controls for UC model validation and 92 CD patients and 108 controls for CD model validation. The Australian validation cohort includes 98 CD patients and 81 controls. The detailed characteristics of discover cohort and validation cohorts were described in the Supplementary Table S1 and Supplementary Table S2. Five in-house sequencing cohorts of non-IBD patients with other gastrointestinal disorders, including colorectal adenomas (CA, n=162), colorectal cancer (CRC, n=160), irritable bowel syndrome (diarrhea subtype, IBS-D, n=117), and non-IBD patients with non-gastrointestinal disorders, including obesity (n=148) and cardiovascular disease (CVD, n=143), were also included for validation. The public datasets of IBD cohorts (n=3,252) and non-IBD cohorts (n=843) were downloaded for validation. Another two independent IBD cohorts from Canada (n=253) and Taiwan (n=120) were used for multiplex droplet digital PCR test and fecal calprotectin test.
Recruitment	IBD patients in Hong Kong discovery cohort and validation cohorts were recruited from Prince of Wales Hospital and several other regional hospitals in the New Territories East Cluster (NTEC) in Hong Kong. Patients with CD and UC were diagnosed according to standard criteria of endoscopy, radiology, and histology. Control subjects were only recruited from Prince of Wales Hospital. Individuals with no existing gut disorders such as inflammatory bowel diseases, cancer, advanced adenoma, irritable bowel syndrome, or other GI symptoms were recruited as control subjects. IBD patients and control subjects were excluded if the antibiotics was used in the last 1 month before entering the study, known current sepsis (excluding uncomplicated infections such as influenza), known history of severe organ failure (including decompensated cirrhosis, malignant disease, kidney failure, epilepsy, active serious infection, acquired immunodeficiency syndrome), major bowel surgery in the last 6 months (excluding colonoscopy/ procedure related to perianal disease), presence of an ileostomy or stoma, or current pregnancy. Finally, 205 UC patients, 174 CD patients and 118 control subjects were recruited in the discovery cohort; 139 UC patients and 139 control subjects were recruited in the UC validation cohort; 92 CD patients and 108 control subjects were recruited in the CD validation cohort. International IBD independent validation cohorts were recruited from Australia, Canada and Taiwan according to the same inclusion and exclusion criteria mentioned above. CD patients and control subjects in Australian validation cohort were recruited from St Vincent's Hospital, Melbourne, Australia. UC patients, CD patients, and control subjects in Canada and Taiwan cohorts were recruited from several centers in Canada and from National Taiwan University Hospital, respectively. Finally, 98 CD patients and 81 control subjects from Australia, 100 UC patients, 100 CD patients and 53 control subjects from Canada, 40 UC patients, 40 CD patients and 40 control subjects from Taiwan were recruited. All non-IBD subjects were recruited from the Prince of Wales Hospital. Subjects with CRC and CA were diagnosed by colonoscopy and confirmed on histology examinations. Subjects with IBS were diagnosed according to the ROME III criteria, and endoscopy and enteroscopy were performed to exclude other GI disorders like IBD, coeliac disease, parasite infestations, and other organic disorders. Subjects with cardiovascular disease (CVD) were recruited from the public as part of a survey of cardiovascular health in the Hong Kong general population. Subjects underwent carotid ultrasounds to measure intima-media thickness (IMT) of the common, internal, and external carotid arteries (CCA, ICA and ECA, respectively) and carotid bulbs and

	subjects that had ≥50% stenosis in a single or multiple vessels were regarded as having the risk of CVD. Patients with obesity were defined as those with body mass index greater than 28.
Ethics oversight	This study has been approved by the Joint Chinese University of Hong Kong - New Territories East Cluster Clinical Research Ethics Committee (The Joint CUHK-NTEC CREC, CRE Ref. No 2013.093, No 2017.495, No 2022.308), the ethics committee of St Vincent's Hospital (Melbourne, Australia), research ethics boards for each center involved in the research in Canada, and National Taiwan University Hospital. All subjects have provided written informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size calculation was not performed before this study because it is a retrospective study. We referenced other biomarker-related publication and comprehensively examined the eligible samples from our in-house biobank. Franzosa, E.A., et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol 4, 293-305 (2019). Ning, L., et al. Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts. Nat Commun 14, 7135 (2023).
Data exclusions	Samples of low read count after quality filtration (<1 million reads) were excluded from our analysis.
Replication	This is a study focusing on the disease diagnosis, and indeed on the reproducibility. We used five-fold cross validation for model construction and validated the models in two independent cohorts and three public datasets from different regions. Droplet digital PCR test were used to validate the robustness of these biomarkers.
Randomization	Not applicable for this observational case-control study.
Blinding	Blinding was impossible during data collection because study subjects were already diagnosed before recruitment. Additionally, blinding was impossible during analysis as models and statistical analyses relied on disease grouping information.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*