Determine independent gut microbiota-diseases association by eliminating the effects of human lifestyle factors

Xin Wang^{a*}, Yuqing Yang^{c*}, Jianchu Li^a, Rui Jiang^d, Ting Chen^{b,#}, Congmin Zhu^{b,#}

^aDepartment of Ultrasound, Peking Union Medical College Hospital, Beijing, China

^bDepartment of Computer Science and Technology and Institute of Artificial Intelligence, Tsinghua University, Beijing, China

^cSogou Inc., Beijing, China

^dBioinformatics Division and Center for Synthetic & Systems Biology, Beijing National Research Center for Information Science and Technology; Department of Automation, Tsinghua University, Beijing 100084, China

11

Running Title: Detecting independent gut microbe-disease association

Address correspondence to Congmin Zhu, congminz@163.com.

14

* Xin Wang and Yuqing Yang contributed equally to this work. Author order was determined on the basis of seniority.

17

ABSTRACT

Human lifestyle and physiological variables on human disease risk have been revealed to be mediated by gut microbiota. Low concordance between many case-control studies for detecting disease-associated microbe existed and it is likely due to the limited sample size and the population-wide bias in human lifestyle and physiological variables. To infer association between whole gut microbiota and diseases accurately, we propose to build machine learning models by including both human variables and gut microbiota based on the American Gut Project data, the largest known publicly available human gut bacterial microbiota dataset. When the model's performance with both gut microbiota and human variables is better than the model with just human variables, the independent association of gut microbiota with the disease will be confirmed. We found that gut microbes showed different association strengths with different diseases. Adding gut microbiota into human variables enhanced the association strengths with inflammatory bowel disease (IBD) and unhealthy status; showed no effect on association strengths with Diabetes and IBS; reduced the association strengths with small intestinal bacterial overgrowth, *C. difficile* infection, lactose intolerance, cardiovascular disease and mental disorders. Our results suggested that although gut microbiota was reported to be associated with many diseases, a considerable proportion of these associations may be spurious. We also proposed a list of microbes as biomarkers to classify IBD and unhealthy status, and validated them by reference to previously published research.

IMPORTANCE we reexamined the association between gut microbiota and multiple diseases via machine learning models on a large-scale dataset, and by considering the effect of human variables ignored by previous studies, truly independent

microbiota-disease associations were estimated. We found gut microbiota is associated independently with IBD and overall health of human, but more evidence is needed to judge associations between microbiota and other diseases. Further functional investigations of our reported disease-related microbes will improve understanding of the molecular mechanism of human diseases.

Keywords: gut microbiota, human variables, disease classification, machine learning

INTRODUCTION

The human intestines are home to a dense microbial community, collectively known as the gut microbiota (1). The gut microbiota forms a complex ecosystem and performs a wide range of functions with far-reaching impacts on human health, including extracting energy from the digestive system, preventing colonization by pathogens, promoting immune homeostasis, producing important metabolites, and even communicating with the central nervous system via the gut-brain axis (2). So, it is thought to play an important role in the development of many diseases, including inflammatory bowel disease (3), *Clostridium difficile* infection (4), Diabetes (5), cardiovascular disease (6), and mental health disorders (7). The gut microbiota determines certain host characteristics and responds to host variables, such as human lifestyle and physiological variables, which can be reflected in the microbial composition (8). Therefore, a considerable part of the human variables on human health and disease risk may be mediated or modified by gut microbiota.

With the development of high-throughput sequencing technology, we are now able to sequence the hypervariable regions of the 16S rRNA gene and cluster into operational taxonomic units (OTUs) to profile the taxonomic composition of the microbial community in an environmental sample (9). Over the last few years, many case-control studies have

been conducted to collect microbial 16S rRNA gene datasets from human fecal samples to explore the associations among the gut microbial community and human diseases to reveal disease-specific microbial biomarkers (10, 11). However, many investigations showed low concordance on the discovered disease-associated microbes and one obvious example is studies about obesity. Gut microbiota reported by multiple studies of which abundance is differential between obese and lean individuals is inconsistent (12). Furthermore, M. A. Sze and P. D. Schloss (13) comprehensively analyzed the results of several obesity-related studies. They found that the statistical detection power of a small-sample study was insufficient, and the ratio of abundance of *Bacteroidetes* and *Firmicutes* was not associated with obesity. In addition, a recent construction of a large dataset from the Swedish population did not reveal an apparent microbial signature associated with irritable bowel syndrome (IBS) as previously reported in the literature, and the heterogeneity of the microbial community among IBS patients was higher than that among healthy individuals (14).

There are two possible reasons which may result in low concordance in previous studies. One is the limitation of the sample size. Generally, there are thousands of microorganisms with a wide range of abundance levels in intestinal samples. Due to the high cost of building a large-scale dataset consisting of both gut microbiota information and elaborate human variables (15), researchers can only afford to sequence dozens or hundreds of samples to explore disease-associated microbes via statistical models. Thus, the model overfitting is common and thus reduces the reliability of the inferred results. The other critical shortcoming is neglecting the influence of host variables, which makes it difficult for researchers to confirm whether the calculated gut microbial-disease

associations indicate the true interactions between microbes and the progression of diseases. The alternative possibility is that microbes are only related to certain host variables, and as a result, they are associated indirectly with diseases(16).

Therefore, a large-scale dataset containing information on both gut microbial community and host variables is required for accurate identification of microbiota-disease associations. Fortunately, the American Gut Project (AGP), which comprised thousands of 16s rRNA gene sequencing samples and a rich human variables set related to human lifestyle and physiological variables and diseases, has been carried out worldwide (17). Today, the AGP has sequenced more than 15,000 samples, which significantly expands human gut microbiota's existing data. Most importantly, it provides a rich resource for each sample with information on gut microbiota, human lifestyle factors, and diseases. The goal of this study is to explore the relationship between these entities using this dataset.

Our approach is different from traditional association inference analysis, which tries to estimate the relationship between a single microbe and a disease. We focus on determining whether the whole gut microbiota is independently associated with human diseases by eliminating the influence of lifestyle factors using machine learning methods (ML). The strength of association between gut microbiota and disease is evaluated by the classification performance of the ML models built with the microbiota. Although researchers have built large microbial datasets by merging different studies to explore the effect of gut microbiota on predicting diseases and mortality risks (18-20) via ML approaches, they neglected the impact of human lifestyle factors, resultant in the magnified predictive power of the microbiota. It is because human lifestyle factors

influence both the gut microbiota and the disease progression. Besides, human lifestyle factors between enrolled healthy individuals (controls) and patients (cases) can be significantly different, and such differences could become the main contributor to the predictive power of the disease. It is not surprising to build a well-performed disease classification model using both gut microbiota and human lifestyle and physiological variables in this condition. However, we argue that the independent gut microbiota-disease associations are real, only when the models' classification performance with both gut microbiota and human variables is significantly better than the model built with just human lifestyle factors (Fig. 1). Conversely, when the models' classification performance is inferior to the human-lifestyle-built model, either gut microbiota may not be associated with diseases, or a more suitable data enrollment criterion is needed.

Following the argument, we explored the classification power of the gut microbiota and human variables on multiple diseases using the AGP data with ML classification models. Key OTUs and human variables, consisting of lifestyle and dietary factors, were identified with high validity using multiple machine learning methods (ML). The performance of OTUs and human variables was compared comprehensively to show the difference between their contributions to diseases. In addition, considering the widespread associations of gut microbes with multiple diseases, we use OTUs to judge the overall health status of human, and individuals with at least one disease were classified as the unhealthy. Although lots of associations with diseases were identified previously, our results showed that adding gut microbiota into human variables only enhanced the association strengths with inflammatory bowel disease (IBD) and unhealthy status. In

addition, we reported the top 10 features (OTUs or phenotypes) used in the classification of these diseases, most of which were supported by previously published studies.

RESULTS

Characteristics of the dataset

Dataset used in this study consisted of 7,571 samples with 517 OTUs and 30 human variables (See MATERIALS AND METHODS for details). For human variables, there were 6 variables related to individuals' physiological characteristics (age, sex, height, weight, body mass index (BMI) and race), 2 related to lifestyle choices (exercise and smoking frequencies), 3 related to location (latitude, elevation and country), and 19 related to diet (frequencies of fruit, high-fat red meat, alcohol, and so on). For every sample, labels of eight diseases [cardiovascular disease (CD), small intestinal bacterial overgrowth (SIBO), mental disorders (MD), lactose intolerance (LI), diabetes (DI), IBD, irritable bowel syndrome (IBS), *C. difficile* infection (CDI) and Diabetes (DI)] that have been reported to be related to gut microbiota were extracted. Besides, a disease label named 'unhealthy (UH)' was added if a sample had at least one of eight diseases. The characteristics of the dataset, the demographic details of samples and the number of male and female patients for each disease are shown in Table S1 and S2.

For every disease, ML classification models were constructed using five types of features respectively: human variables only (Meta), OTU abundance only (OTUab), OTU occurrence only (OTUoc), both human variable data and OTU abundance (Meta-OTUab), and both human variable data and OTU occurrence (Meta-OTUoc). Models were trained and compared on identical training and validation data. For each type of feature, the best model was selected according to the AUC score. By comparing the performance of

models with only Meta and models with both human variables and OTU information (Meta-OTUab and Meta-OTUoc), all diseases were classified into three categories: adding gut microbiota a) could improve, b) didn't affect or c) reduced disease classification performance.

Adding gut microbiota into human variables enhanced the association strengths with IBDAs a global public health concern, the incidence and prevalence of IBD, which is caused by gut dysbiosis, is increasing in developed and developing countries (21, 22). In this study, 413 IBD patients from the AGP were included in the final dataset, and the results of the best models using five types of features are shown in Table 1. The model using human variable data only (Meta) as the feature achieved an AUC of 0.72680 ± 0.00874 . Interestingly, the AUCs of using OTUab alone (0.71404 ± 0.01229) and OTUoc alone (0.70693 ± 0.01536) did not differ significantly from that obtained using Meta ($P = 0.0096$), indicating that gut microbiota alone is as good a classifier of IBD as human variables. Additionally, the AUCs obtained using the combination of Meta and OTUs (Meta-OTUab: 0.74706 ± 0.01421 and Meta-OTUoc: 0.75464 ± 0.01855) were significantly higher than those obtained using Meta alone ($P = 0.00037$ and $P = 0.00057$, respectively), suggesting that adding gut microbiota into human variables significantly enhanced the association strengths with IBD (Table 1) and that the independent association of gut microbiota with IBD can be confirmed. It is noteworthy that the Meta-OTUoc achieved higher AUCs than those achieved with Meta-OTUab, which implied that, compared to the abundance of gut microbes, their occurrences are better features for the classification of IBD.

Next, we assessed the relative roles of Meta or OTUs in the best ML model for classifying IBD with Meta, Meta-OTUab, and Meta-OTUoc. We ranked the features

according to their weights and calculated the average rank after repeating the model training process ten times. As shown in Table 2, we found that the top 10 features for three types of features were distinct. For the model using human variable data only (Meta) as features, the top 10 most important human variables for classifying IBD comprised six dietary characteristics (the frequencies of vitamin B, alcohol, vitamin D, whole grain, probiotic, and red meat intake) and four basic physical characteristics (BMI, age, height, and weight), without geographical location features. For the model using Meta-OTUab as features, except for three dietary characteristics (the frequencies of vitamin B, vitamin D and probiotic intake), the other seven of the top 10 features for classifying IBD were all OTUs (four *Clostridiales*, one *Bacteroidales*, one *Erysipelotrichales* and one *Pasteurellales*). When using Meta-OTUoc as features to classify IBD, the results changed in that, except for two physical characteristics (age and height) and one OTU of *Enterobacteriales*, the other seven of the top 10 features were all OTUs of *Clostridiales*.

Adding gut microbiota showed no effect on association strengths with Diabetes and IBS

DI and IBS are widely reported to be closely related to gut microbes and some diet habits (3, 23); therefore, we hypothesized that adding microbes into human variables will improve the classification of these two conditions. As shown in Table 2, Meta provided the best performance in classifying both these two diseases. The maximal AUCs for classifying CDI with Meta were 0.84658 ± 0.02231 . The AUCs obtained using OTUab alone (0.66839 ± 0.02136) and OTUoc alone (0.66216 ± 0.03878) were significantly lower than those obtained using Meta. It is noteworthy that the AUCs obtained using the combination of Meta and OTUs (Meta-OTUab: 0.82068 ± 0.02053 and Meta-OTUoc: 0.83591 ± 0.01429) did not differ significantly from those obtained using Meta alone ($P =$

0.0506 and $P = 0.0864$, respectively), suggesting that adding gut microbiota into human variables showed no effect on association strengths with DI. Meta-OTUoc achieved higher AUCs than Meta-OTUab. The ROC curves of the models based on Meta-OTUab and Meta-OTUoc were close to those of the Meta-only based models (Fig. 2). According to the feature weights, we identified the top 10 most important human variables for classifying DI (Fig. 3 and Table S3). When using Meta as the feature set, the top 10 most important human variables for IBS comprised four basic physical characteristics (BMI, age, weight and height), two geographical location features (elevation and latitude), three dietary characteristics (the frequencies of alcohol, milk cheese and milk intake) and one lifestyle habit (exercise frequency). For the model with Meta-OTUab as features, with the exception of two basic physical characteristics (BMI and age), eight of the top 10 most important features for classifying DI were OTUs (one *Desulfovibrionales*, one *Actinomycetales*, four *Clostridiales*, one *Bacteroidales* and one *Lactobacillales*). For the model with Meta-OTUoc as the features for DI, the most important features were BMI and age, followed by two OTUs belonging to *Ruminococcaceae* and six human variables (height, latitude, elevation, the frequency of milk, alcohol, and egg intake). Interestingly, BMI and age were the most two most important features in all three models.

When classifying IBS, the best model using Meta achieved an AUC of 0.68448 ± 0.00780 . The AUCs obtained using OTUab alone (0.59820 ± 0.00995), OTUoc alone (0.56972 ± 0.01460) and Meta-OTUab (0.66473 ± 0.01118) were significantly lower than those obtained using Meta. However, the AUCs obtained using Meta-OTUoc (0.66950 ± 0.01386) did not differ significantly from those obtained using Meta alone (Meta), suggesting that adding gut microbiota occurrence features did not improve the

association strength with IBS. According to the feature weights, the top 10 most important human variables for IBS using Meta and Meta-OTUoc are shown in Table S4. When using Meta as the feature set, the top 10 most important human variables consisted of two geographical location features (latitude and elevation), four basic physical characteristics (age, weight, BMI and height), three dietary characteristics (the frequencies of probiotic, cheese and seafood intake), and one lifestyle habit (exercise frequency). For the model using Meta-OTUoc as the features, with the exception of one OTU (*Enterobacter radicincitans*), the other nine of the top 10 most important features were all human variables, comprising four dietary characteristics (the frequencies of milk cheese, probiotic, whole grain, and seafood intake), one geographical location features (latitude) and four basic physical characteristics (age, weight, sex, and height).

Adding gut microbiota reduced the association strengths with the other five diseases

Recently, gut microbes were also reported to be related to small intestinal bacterial overgrowth (SIBO), *C. difficile* infection (CDI), lactose intolerance (LI), cardiovascular disease (CD) and mental disorders (MD) (4-7, 24). As shown in Table 1, for all five diseases, the models using Meta as the features achieved the highest AUC values (SIBO: 0.74413 ± 0.0234 , CDI: 0.84091 ± 0.01890 , LI: 0.81647 ± 0.00751 , CD: 0.83432 ± 0.01255 , MD: 0.62984 ± 0.01240). Using OTUs alone (OTUab and OTUoc) or the OTUs combined with human variables (Meta-OTUab and Meta-OTUoc) resulted in AUCs that were significantly lower than those obtained using the human variables alone (Meta). Although many gut microbes are reported to be associated with these diseases, adding gut microbiota reduced the association strengths with the other five diseases (Table 1 and Fig. 2). In addition, neither Meta nor OTUs provide good performance for classifying MD.

The top 10 most important human variables for classifying SIBO, CDI, LI, CD and MD were identified according to the feature weights (Fig. 3 and Table S5). The top 10 most important human variables for classifying SIBO comprised five dietary characteristics (the frequencies of whole grain, milk cheese, probiotic, frozen dessert and poultry intake), four basic physical characteristics (weight, BMI, height and age) and one geographical location feature (latitude). With the exception of three basic physical characteristics (height, BMI and age) and two geographical location features (latitude and elevation), five of the top 10 most important human variables for classifying CDI were related to dietary characteristics (the frequencies of probiotic, vitamin B, alcohol, whole eggs and cheese intake). It is reasonable that the two most important human variables for classifying LI are the frequencies of milk cheese and milk substitute intake and age, followed by one race-related feature (Caucasian), five dietary characteristics (the frequencies of alcohol, poultry, high-fat red meat, salted snacks and probiotic intake), height and elevation. The top 10 most important human variables for classifying CD comprised five basic physical characteristics (age, sex, height, weight and BMI), three dietary characteristics (the frequencies of high-fat red meat, olive oil and fruit intake) and two geographical location features (latitude and elevation). It is noteworthy that MD was classified mainly by four basic physical characteristics (age, BMI, height and weight), four dietary characteristics (the frequencies of milk, alcohol, vitamin D and fermented plant intake) and two geographical location features (latitude and elevation).

Adding gut microbiota improved the association strength with unhealthy status

To investigate the potential use of the out composition to classify the health of individuals, we defined a sample as unhealthy if it was obtained from an individual with

any one of the eight diseases. Finally, we obtained 2,921 unhealthy samples containing at least one of those eight diseases for training the models using OTUs or Meta alone and in combination. As shown in Table 1, the model using Meta as the features achieved an AUC of 0.70936 ± 0.00442 . The AUCs obtained using Meta-OTUab (0.75626 ± 0.00859) and Meta-OTUoc (0.76266 ± 0.00520) were all significantly higher than those obtained using Meta, suggesting that adding gut microbiota improved the accuracy of classifying unhealthy status. Surprisingly, the AUC obtained using OTUab (0.71976 ± 0.00533) alone was also significantly higher than that obtained using Meta alone, indicating that unhealthy status can be classified accurately based on the abundance information of gut microbes alone. However, the AUC obtained using OTUoc (0.71976 ± 0.00533) was not significantly different from that obtained using Meta. When calculating the weight for each feature in the models with Meta, OTUab, Meta-OTUab and Meta-OTUoc, we identified the top 10 most important features for classifying unhealthy status (Table 3). For the model with Meta as the features, the top 10 most important human variables comprised five dietary characteristics (the frequencies of cheese, probiotic, milk substitute, vitamin D and frozen dessert intake), three basic physical characteristics (age, BMI and weight) and two geographical location features (latitude and elevation). For the model using OTUab as the features, eight of the top 10 OTUs were annotated to *Clostridiales*, one to *Bacteroidales* and one to *Bifidobacteriales*. When using Meta-OTUab as the features, with the exception of four dietary characteristics (the frequencies of milk cheese, probiotic, vitamin D and milk substitute intake) and age, the other five of the top 10 most important features were all OTUs (four annotated to *Clostridiales* and one *Bacteroidales*). When using Meta-OTUoc as the features to classify unhealthy status, the top 10 most important

features comprised five dietary characteristics (the frequencies of milk cheese, probiotics, vitamin D, milk substitute and frozen dessert intake), one basic physical characteristic (age), one geographical location feature (latitude) and three OTUs annotated to *Clostridiales*.

DISCUSSION

Important OTUs for human diseases

For IBD, adding gut microbiota to human variables can achieve better results than that achieved using human variables alone and Meta-OTUoc achieved higher AUCs than that using Meta-OTUab. Among the top 10 most important features used to classify IBD with Meta-OTUoc, eight were OTUs with one belonging to the *Klebsiella* genus. *Klebsiella* is an intestinal pathobiont that can produce a cytotoxin (tillivalline), and is thought to be involved in the pathogenesis of IBD (25). It is also interesting that seven of the eight OTUs belonged to the *Clostridiales* order. At the family level, two were annotated to *Ruminococcaceae*, which have been reported as a prominent family in IBD, especially two species, *Ruminococcus torques* and *Ruminococcus gnavus* (26) and three were annotated to *Lachnospiraceae*, which is also reported to be related to IBD (27). The physiologic niche of *Ruminococcus gnavus* was speculated to be mucolytic, with dramatic changes in this species affecting the delicate equilibrium of the mucus layer and potentially increasing the intestinal permeability in IBD patients (26). For the unhealthy status classification, the AUC obtained using OTUab alone was significantly higher than that obtained using Meta alone, which shows that unhealthy status can be classified accurately using only the abundance information of gut microbes, and that gut microbes may be good biomarkers for the general health condition of humans. Eight of the top 10 OTUs with the highest

weight belonged to the *Clostridiales* order. Under the *Clostridiales* order, the *Lachnospiraceae* family has been associated with many human diseases, such as inflammatory bowel disease (27), irritable bowel syndrome (28), type 1 diabetes (29, 30), *Clostridium difficile* infection (31) and liver cirrhosis (32). *Ruminococcaceae* family was also reported to be related to *Clostridium difficile* infection (31) and type 1 diabetes (29). The other two most important OTUs for unhealthy status classification belonged to the *Bacteroides* genus and the *Bifidobacterium* genus. According to the previous research, *Bacteroides* genus was associated with several human diseases, including five gut diseases (irritable bowel syndrome (28), *Clostridium difficile* infection (31), colorectal carcinoma (33), Crohn's disease (34, 35) and infectious colitis (36)), type 1 diabetes (29, 30) and liver cirrhosis (37).

Important human variables for human diseases

Human variables showed a strong efficiency in human disease classification. According to our results shown in Fig. 3, the basic physiological characteristics (BMI, age, height and weight) are the most important human variables correlated to most diseases, followed by location factors (latitude and elevation) and the frequency of probiotic, milk cheese and alcohol intake. BMI and age were found to be important classifiers of all seven diseases except lactose intolerance, which is supported by a cross-sectional study of pre-specified demographic and clinical data (38). Multiple pieces of evidence from experimental and observational studies showed that for a substantial proportion of patients with IBS, their symptoms were associated with the ingestion of specific foods, such as milk, which contain lactose, a disaccharide that is not effectively digested by many adults worldwide (39). Additionally, evidence exists to suggest that probiotics may exert an effect on IBS

through various mechanisms (40). In accordance with previous studies, we found that milk cheese and probiotic intake were two of the most important types of human variables in addition to attitude and age for classifying IBS (Table S4). For DI classification, we found that BMI and age were the two most important types of human variables, which was supported by a cross-sectional study (38). We found that the most important health feature for classifying SIBO was the frequency of probiotic intake, which was supported by a previous systematic review (41). The review and meta-analysis showed that probiotics are both safe and effective for preventing SIBO. We also found that the frequency of cheese and milk intake are two of the most important features for LI classification. This finding is not surprising because the breakdown of nondigested lactose causes LI; therefore, LI management usually involves excluding milk and milk products from the diet. It is noteworthy that the race-related feature was also among the top 10 health features for LI classification. This discovery was validated by the previous reports that lactase persistence varies among different human populations (42). Age and sex were found to be the two most important features for classifying CD. This is a reasonable conclusion because age has been reported as one of the most powerful risk factors for developing CD (43) and there is a higher prevalence of CD in men than in women (44).

Removing probiotics, vitamin B and vitamin D

The frequency of probiotic intake was among the top 10 features of IBD, CDI, IBS, SIBO, and LI; the frequency of eating vitamin B was among the top ten features of CDI and IBD; the frequency of vitamin D intake was among the top 10 features of IBD and MD (Fig. 3). Some samples may have been obtained from individuals who adopted these three dietary habits advised by the clinician; therefore, we repeated our analysis after removing these

three health features. As the results are shown in Table 4, we found that, after removing the frequency of probiotic and vitamins B and D intake, Meta-OTUoc showed significantly better results than Meta for both classifying IBD and unhealthy status, showing that the independent associations do exist between gut microbiota and IBD and unhealthy status. It is interesting that in addition to DI and IBS, Meta-OTUoc also did not differ significantly from Meta for the classification of SIBO and MD. For the other three diseases (CDI, LI and CD), Meta still showed the best performance. The top 10 most important features for disease classification after removing the frequency of probiotic and vitamins B and D intake are shown in Table S6.

Model performance changed with OTU numbers

When there are too many OTUs as input features, the models may be overfitted. Therefore, we evaluated the changes in classification performance for the four diseases (IBD, DI, IBS and UH) in the optimal model result with the number of OTUs. The changes in AUCs obtained using the number of OTUs are shown in Fig. 4. We found that using only some of the OTUs achieved better results than using all 517 OTUs. Especially for DI, using only the top 1% of OTUs (10 OTUs) achieved the best effect. Furthermore, the top 3% of OTUs (20 OTUs) generated the best unhealthy status and IBS classification results with AUCs of 0.766 and 0.701, while the top 77% of OTUs (400 OTUs) generated the best IBD classification results with an AUC of 0.761. The OTU set with the best classification results were different for the four diseases. These OTU sets can be used as biomarkers for the corresponding diseases. In addition, when we used only gut microbes for disease classification, the OTUab achieved better results than that achieved using OTUoc, which can be explained by the greater loss of information using OTUoc than

when using OTUab. However, after combination with human variables, the performance of Meta-OTUoc surpassed the Meta-OTUab in most cases. This difference might be caused by the reason that values of human variable data were associated with microbial abundance, and, for disease classification, information provided by human variables and microbial abundance were overlapped.

MATERIALS AND METHODS

Data sources

We downloaded the OTU table (11-packaged/fecal/100nt/ag_fecal.biom) and human variables (11-packaged/fecal/100nt/ag_fecal.txt) from the latest version (updated in January 2018) of the AGP database available at . The original OTU table was saved as a binary file (.biom), which was converted manually to plain text with Python Script, which is available at GitHub (<https://github.com/tinglab/kLDM.git>). The original gut microbial abundance table (OTU table) contained 15,158 samples and 24,114 OTUs selected by applying a 97% similarity cutoff with SortMeRNA defined by the AGP consortium. The OTUs were mapped to the Greengenes Database (45) to identify their taxonomy. Each cell in the OTU table presents the abundance of its corresponding OTUs in a specific sample. The original human variables file (Meta table) contained 15,158 samples and 523 factors related to physicochemical parameters of fecal samples, dietary habits, lifestyle choices and some diseases. Each cell in the Meta table presents the measured value of its corresponding meta-data in a specific sample.

Data preprocessing

We selected 30 items of human variables to classify disease, including individuals' physiological characteristics, lifestyle, location, and diet. Among these 30 items, six were

related to physiological characteristics; two were associated with lifestyle choices; three were associated with location; the remaining 19 were connected with diet (frequencies of fermented plant, frozen dessert, fruit, high-fat red meat, home-cooked meals, alcohol, red meat, meat eggs, milk substitute, milk cheese, olive oil, probiotic, salted snacks, seafood, vegetable, vitamin D supplement, vitamin B supplement, whole grain and whole eggs). The values of frequency-related human variables were categorized as follows: 'Never', 'Rarely' (less than once/week), 'Occasionally' (1–2 times/week), 'Regularly' (3–5 times/week) and 'Daily'. For convenience, these categories were recoded as integers from 1 to 5 (where 1 represents 'never' and 5 represents 'daily') according to their frequencies (Table S1). The values of human variables were missing for some samples; therefore, the samples with complete sets of 30 human variable data items were selected for the following analysis. Samples with huge (first 1%) and small (last 2%) reads, as well as those with evenness <2 were removed. For all selected samples, OTUs were filtered based on their average size and non-zero times. Finally, 7,571 samples with 517 OTUs and complete human variables were obtained.

Eight diseases (CD, SIBO, MD, LI, DI, IBD, IBS, CDI and DI) that have been reported in previous studies to be related to gut microbiome were selected. The samples from individuals affected by any of these eight diseases were treated as unhealthy (UH). Information about individuals' disease status was extracted, and the samples from individuals with diseases were labeled. The characteristics of the dataset, the demographic details of samples and the number of male and female patients for each disease are shown in Table S2.

Machine learning models training and evaluation

To evaluate the confound association between disease with gut microbiota and human variables, four machine learning (ML) techniques (RF, Random Forest; GBDT, Gradient Boosting Decision Tree; LR, Logistic Regression; XGBoost, eXtreme Gradient Boosting) were used to build the model, and the AUC scores were calculated to compare their performance (Fig. 1). For every disease, four types of ML models were trained with five-fold cross-validation using training data, including 90% of all samples, and the model providing the best performance was selected based on the maximal AUC. Considering that the positive samples labeled with diseases occupied a tiny proportion, an equal number of negative samples were randomly selected for model training. The optimal model was then evaluated and compared using the validation dataset comprising the remaining 10% of samples.

In addition to the different model types, five combinations of features were used to construct the following separate models to capture the best features for classifying each disease: human variables only (Meta), OTU abundance only (OTUab), OTU occurrence only (OTUoc), both human variable data and OTU abundance (Meta-OTUab), and both human variable data and OTU occurrence (Meta-OTUoc). The OTU occurrence was determined based on the existence of OTUs only. Models using different combinations of features were trained and compared using an identical dataset. For each type of feature, the best model with the maximal AUC score was selected from the four types of models.

To assess the significance of differences in the model performance among the five types of features, the model training process was repeated 10 times with a random selection of training data. The AUC scores were presented as the mean $\pm$ standard deviation. For each

disease, paired sample *t*-tests were used to compare the differences in AUC values between the feature type 'Meta' and the four other feature types ('OTUab', 'OTUoc', 'Meta-OTUab' and 'Meta-OTUoc'). In the statistical analysis, Bonferroni correction was used to adjust for the multiple testing error. In considering nine disease and four comparisons, we tested 36 independent hypotheses using the same data at the 0.05 significance level, and instead of using a *P*-value threshold of 0.05, we use a stricter threshold of 0.0014.

Identification of microbial biomarkers of diseases

For the best performing model of each disease, the weights of each feature were calculated. The top ten features (OTUs or Meta) with the highest absolute weights were selected as the biomarkers for the disease. We then obtained the taxa of these OTUs and verified their relationships to the disease by searching published literature or databases. In this study, we used the Human Microbe-Disease Association Database (HMDAD) (<http://www.cuilab.cn/hmdad>), which is a curated collection of microbe-disease association from previous microbiota studies. The OTUs with high weights that were verified were treated as microbial biomarkers of the disease.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (Grant NO. 61872218 and 61721003), China Postdoctoral Science Foundation (Grant NO. 2020M680454), Tsinghua University Students Research Training (Grant NO. 53412002520) and Beijing National Research Center for Information Science and Technology (BNRist). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

All authors conceived the study and participated in its design. CMZ, XW performed the collecting and cleaning of datasets. XW, YQY analyzed the data and performed the model analysis. CMZ, JCL, RJ and TC performed manuscript preparation. All authors read and approved the final manuscript.

REFERENCES

1. Suau A, Bonnet R, Sutren M, Godon JJ, Gibson GR, Collins MD, Doré J. 1999. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* 65:4799-807.
2. Foster JA, McVey Neufeld KA. 2013. Gut-brain axis: how the microbiome influences anxiety and depression. *Trends Neurosci* 36:305-12.
3. Chu H, Khosravi A, Kusumawardhani IP, Kwon AH, Vasconcelos AC, Cunha LD, Mayer AE, Shen Y, Wu WL, Kambal A, Targan SR, Xavier RJ, Ernst PB, Green DR, McGovern DP, Virgin HW, Mazmanian SK. 2016. Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. *Science* 352:1116-20.
4. Samarkos M, Mastrogianni E, Kampouroupoulou O. 2018. The role of gut microbiota in *Clostridium difficile* infection. *Eur J Intern Med* 50:28-32.
5. Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T, Jensen BA, Forslund K, Hildebrand F, Prifti E, Falony G, Le Chatelier E, Levenez F, Doré J, Mattila I, Plichta DR, Pöhö P, Hellgren LI, Arumugam M, Sunagawa S, Vieira-Silva S, Jørgensen T, Holm JB, Trošt K, Kristiansen K, Brix S, Raes J, Wang J, Hansen T, Bork P, Brunak S, Oresic M, Ehrlich SD, Pedersen O. 2016. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* 535:376-81.
6. Jie Z, Xia H, Zhong SL, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H, Zhang D, Su Z, Fang Z, Lan Z, Li J, Xiao L, Li J, Li R, Li X, Li F, Ren H, Huang Y, Peng Y, Li G, Wen B, Dong B, Chen JY, Geng QS, Zhang ZW, Yang H, Wang J, Wang J, Zhang X, Madsen L, Brix S, Ning G, Xu X, Liu X, Hou Y, Jia H, He K, Kristiansen K. 2017. The gut microbiome in atherosclerotic cardiovascular disease. *Nat Commun* 8:845.
7. Sampson TR, Mazmanian SK. 2015. Control of brain development, function, and behavior by the microbiome. *Cell Host Microbe* 17:565-76.
8. Fan Y, Pedersen O. 2020. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol* doi:10.1038/s41579-020-0433-9.
9. Hao X, Jiang R, Chen T. 2011. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 27:611-8.
10. Oozeer R, Rescigno M, Ross RP, Knol J, Blaut M, Khlebnikov A, Doré J. 2010. Gut health: predictive biomarkers for preventive medicine and development of functional foods. *British journal of nutrition* 103:1539-1544.
11. Tackmann J, Arora N, Schmidt TSB, Rodrigues JFM, von Mering C. 2018. Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites. *Microbiome* 6:1-16.

- 516 12. Stanislawski MA, Dabelea D, Lange LA, Wagner BD, Lozupone CA. 2019. Gut microbiota
517 phenotypes of obesity. *NPJ biofilms and microbiomes* 5:1-9.
- 518 13. Sze MA, Schloss PD. 2016. Looking for a signal in the noise: revisiting obesity and the microbiome.
519 *MBio* 7:e01018-16.
- 520 14. Hugerth LW, Andreasson A, Talley NJ, Forsberg AM, Kjellström L, Schmidt PT, Agreus L,
521 Engstrand L. 2020. No distinct microbiome signature of irritable bowel syndrome found in a
522 Swedish random population. *Gut* 69:1076-1084.
- 523 15. Manor O, Dai CL, Kornilov SA, Smith B, Price ND, Lovejoy JC, Gibbons SM, Magis AT. 2020.
524 Health and disease markers correlate with gut microbiome composition across thousands of people.
525 *Nature communications* 11:1-12.
- 526 16. Yang Y, Chen N, Chen T. 2017. Inference of environmental factor-microbe and microbe-microbe
527 associations from metagenomic data using a hierarchical Bayesian statistical model. *Cell systems*
528 4:129-137. e5.
- 529 17. McDonald D, Birmingham A, Knight R. 2015. Context and the human microbiome. *Microbiome*
530 3:1-8.
- 531 18. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of
532 large metagenomic datasets: tools and biological insights. *PLoS computational biology*
533 12:e1004977.
- 534 19. Tierney BT, He Y, Church GM, Segal EJ, Kostic AD, Patel CJ. 2020. The predictive power of the
535 microbiome exceeds that of genome-wide association studies in the discrimination of complex
536 human disease. *bioRxiv*:2019.12. 31.891978.
- 537 20. Salosensaari A, Laitinen V, Havulinna AS, Meric G, Cheng S, Perola M, Valsta L, Alfthan G,
538 Inouye M, Watrous JD. 2020. Taxonomic Signatures of Long-Term Mortality Risk in Human Gut
539 Microbiota. *medRxiv*:2019.12. 30.19015842.
- 540 21. Khan I, Ullah N, Zha L, Bai Y, Khan A, Zhao T, Che T, Zhang C. 2019. Alteration of gut
541 microbiota in inflammatory bowel disease (IBD): Cause or consequence? IBD treatment targeting
542 the gut microbiome. *Pathogens* 8:126.
- 543 22. Knox NC, Forbes JD, Van Domselaar G, Bernstein CN. 2019. The gut microbiome as a target for
544 IBD treatment: are we there yet? *Current treatment options in gastroenterology* 17:115-126.
- 545 23. Labus JS, Hollister EB, Jacobs J, Kirbach K, Oezguen N, Gupta A, Acosta J, Luna RA, Aagaard K,
546 Versalovic J, Savidge T, Hsiao E, Tillisch K, Mayer EA. 2017. Differences in gut microbial
547 composition correlate with regional brain volumes in irritable bowel syndrome. *Microbiome* 5:49.
- 548 24. Pylaris E, Giamarellos-Bourboulis EJ, Tzivras D, Koussoulas V, Barbatzas C, Pimentel M. 2012.
549 The prevalence of overgrowth by aerobic bacteria in the small intestine by small bowel culture:
550 relationship with irritable bowel syndrome. *Dig Dis Sci* 57:1321-9.
- 551 25. Zollner-Schwetz I, Herzog KA, Feierl G, Leitner E, Schneditz G, Sprenger H, Prattes J, Petritsch W,
552 Wenzl H, Kump P, Gorkiewicz G, Zechner E, Högenauer C. 2015. The Toxin-Producing
553 Pathobiont *Klebsiella oxytoca* Is Not Associated with Flares of Inflammatory Bowel Diseases. *Dig*
554 *Dis Sci* 60:3393-8.
- 555 26. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E,
556 Ajami NJ, Bonham KS, Brislawn CJ, Casero D, Courtney H, Gonzalez A, Graeber TG, Hall AB,
557 Lake K, Landers CJ, Mallick H, Plichta DR, Prasad M, Rahnavard G, Sauk J, Shungin D,
558 Vazquez-Baeza Y, White RA, 3rd, Braun J, Denson LA, Jansson JK, Knight R, Kugathasan S,

559 McGovern DPB, Petrosino JF, Stappenbeck TS, Winter HS, Clish CB, Franzosa EA, Vlamakis H,
560 Xavier RJ, Huttenhower C. 2019. Multi-omics of the gut microbial ecosystem in inflammatory
561 bowel diseases. *Nature* 569:655-662.

562 27. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. 2007.
563 Molecular-phylogenetic characterization of microbial community imbalances in human
564 inflammatory bowel diseases. *Proc Natl Acad Sci U S A* 104:13780-5.

565 28. Saulnier DM, Riehle K, Mistretta TA, Diaz MA, Mandal D, Raza S, Weidler EM, Qin X, Coarfa C,
566 Milosavljevic A, Petrosino JF, Highlander S, Gibbs R, Lynch SV, Shulman RJ, Versalovic J. 2011.
567 Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome.
568 *Gastroenterology* 141:1782-91.

569 29. Giongo A, Gano KA, Crabb DB, Mukherjee N, Novelo LL, Casella G, Drew JC, Ilonen J, Knip M,
570 Hyöty H, Veijola R, Simell T, Simell O, Neu J, Wasserfall CH, Schatz D, Atkinson MA, Triplett
571 EW. 2011. Toward defining the autoimmune microbiome for type 1 diabetes. *Isme j* 5:82-91.

572 30. Brown CT, Davis-Richardson AG, Giongo A, Gano KA, Crabb DB, Mukherjee N, Casella G, Drew
573 JC, Ilonen J, Knip M, Hyöty H, Veijola R, Simell T, Simell O, Neu J, Wasserfall CH, Schatz D,
574 Atkinson MA, Triplett EW. 2011. Gut microbiome metagenomics analysis suggests a functional
575 model for the development of autoimmunity for type 1 diabetes. *PLoS One* 6:e25792.

576 31. van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, de Vos WM, Visser CE, Kuijper
577 EJ, Bartelsman JF, Tijssen JG, Speelman P, Dijkgraaf MG, Keller JJ. 2013. Duodenal infusion of
578 donor feces for recurrent *Clostridium difficile*. *N Engl J Med* 368:407-15.

579 32. Chen Y, Yang F, Lu H, Wang B, Chen Y, Lei D, Wang Y, Zhu B, Li L. 2011. Characterization of
580 fecal microbial communities in patients with liver cirrhosis. *Hepatology* 54:562-72.

581 33. Sobhani I, Tap J, Roudot-Thoraval F, Roperch JP, Letulle S, Langella P, Corthier G, Tran Van
582 Nhieu J, Furet JP. 2011. Microbial dysbiosis in colorectal cancer (CRC) patients. *PLoS One*
583 6:e16393.

584 34. Dicksved J, Halfvarson J, Rosenquist M, Järnerot G, Tysk C, Apajalahti J, Engstrand L, Jansson JK.
585 2008. Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *Isme j*
586 2:716-27.

587 35. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C,
588 Chardon P, Marteau P, Roca J, Dore J. 2006. Reduced diversity of faecal microbiota in Crohn's
589 disease revealed by a metagenomic approach. *Gut* 55:205-11.

590 36. Sokol H, Seksik P, Rigottier-Gois L, Lay C, Lepage P, Podglajen I, Marteau P, Doré J. 2006.
591 Specificities of the fecal microbiota in inflammatory bowel disease. *Inflamm Bowel Dis* 12:106-11.

592 37. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, Zhou J, Ni S,
593 Liu L, Pons N, Batto JM, Kennedy SP, Leonard P, Yuan C, Ding W, Chen Y, Hu X, Zheng B, Qian
594 G, Xu W, Ehrlich SD, Zheng S, Li L. 2014. Alterations of the human gut microbiome in liver
595 cirrhosis. *Nature* 513:59-64.

596 38. van Keulen K, Knol W, Schrijver EJM, van Marum RJ, van Strien AM, Nanayakkara PWB. 2018.
597 Prophylactic Use of Haloperidol and Changes in Glucose Levels in Hospitalized Older Patients. *J*
598 *Clin Psychopharmacol* 38:51-54.

599 39. Cozma-Petruț A, Loghin F, Miere D, Dumitrașcu DL. 2017. Diet in irritable bowel syndrome:
600 What to recommend, not what to forbid to patients! *World J Gastroenterol* 23:3771-3783.

601 40. Simrén M, Barbara G, Flint HJ, Spiegel BM, Spiller RC, Vanner S, Verdu EF, Whorwell PJ,
602 Zoetendal EG. 2013. Intestinal microbiota in functional bowel disorders: a Rome foundation report.
603 Gut 62:159-76.

604 41. Simpson M, Lyon C. 2019. PURL: Do probiotics reduce C diff risk in hospitalized patients? J Fam
605 Pract 68:351;352;354.

606 42. Lomer MC, Parkes GC, Sanderson JD. 2008. Review article: lactose intolerance in clinical
607 practice--myths and realities. Aliment Pharmacol Ther 27:93-103.

608 43. Lind L, Ingelsson E, Sundström J, Siegbahn A, Lampa E. 2018. Methylation-based estimated
609 biological age and cardiovascular disease. Eur J Clin Invest 48.

610 44. Spence JD, Pilote L. 2015. Importance of sex and gender in atherosclerosis and cardiovascular
611 disease. Atherosclerosis 241:208-10.

612 45. Desantis TZ, Hugenholtz P, ., Larsen N, ., Rojas M, ., Brodie EL, Keller K, ., Huber T, ., Dalevi
613 D, ., Hu P, ., Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and
614 workbench compatible with ARB. Applied & Environmental Microbiology 72:5069-72.

615

616 **Figures**

617 *Fig.1 Workflow of disease classification models construction. A) We propose to build*
618 *association models by including both human variables and gut microbiota from*
619 *American Gut Project, and assumed that when the performance of the model with both*
620 *gut microbiota and human variables is better than the model with just human variables,*
621 *the independent association of gut microbiota with the disease can be confirmed. B) We*
622 *classified eight diseases (IBD: Inflammatory Bowel Disease; CDI: C. Difficile Infection;*
623 *IBS: Irritable Bowel Syndrome; SIBO: Small Intestinal Bacterial Overgrowth; DI:*
624 *Diabetes; LI: Lactose Intolerance; CD: Cardiovascular Disease; MD: Mental Disorder)*
625 *with human variables (physiological characteristics, lifestyle, location and diet) and gut*
626 *microbial community data (OTUs) obtained from the American Gut Project (AGP)*
627 *database using four machine learning techniques (Random Forest, Gradient Boosting*
628 *Decision Tree, Logistic Regression and eXtreme Gradient Boosting).*

629

630

Fig. 2 ROC curves for selection of the best classification models of eight diseases (IBD: Inflammatory Bowel Disease; CDI: C. Difficile Infection; IBS: Irritable Bowel Syndrome; SIBO: Small Intestinal Bacterial Overgrowth; DI: Diabetes; LI: Lactose Intolerance; CD: Cardiovascular Disease; MD: Mental Disorder). Line colors indicate different input features for the models.

Fig. 3 Top 10 health features for the other seven diseases (IBD: inflammatory bowel disease; CDI: C. Difficile Infection; IBS: Irritable Bowel Syndrome; SIBO: Small Intestinal Bacterial Overgrowth; DI: Diabetes; LI: Lactose Intolerance; CD: Cardiovascular Disease; MD: Mental Disorder). The direct line from one feature to one disease indicates the feature is one of the top 10 health features of the disease. The seven diseases are shown in red. The health features are shown in green, with the intensity of node shading indicating a greater disease classification value of the feature.

Fig. 4 Changes in the AUC of the optimal model with the number of OTUs. The optimal model for four diseases (IBD: Inflammatory Bowel Disease; IBS: Irritable Bowel Syndrome; DI: Diabetes; UH: Unhealthy status).

650

651

652 Tables

653 *Table 1. Comparing AUC values of nine diseases using five feature types.*

Feature type		Meta	OTUab	OTUoc	Meta-OT Uab	Meta-OT Uoc
IBD	AUC	0.72680 ± 0.00874	0.71404 ± 0.01229	0.70693 ± 0.01536	0.74706 ± 0.01421	0.75464 ± 0.01855
	p-value	-	0.00965	0.00256	0.00037	0.00057
DI	AUC	0.84658 ± 0.02231	0.66839 ± 0.02136	0.66216 ± 0.03878	0.82068 ± 0.02053	0.83591 ± 0.01429
	p-value	-	<0.0001	<0.0001	0.05067	0.0864
IBS	AUC	0.68448 ± 0.00780	0.59820 ± 0.00995	0.56972 ± 0.01460	0.66473 ± 0.01118	0.66950 ± 0.01386
	p-value	-	<0.0001	<0.0001	0.00107	0.00967
SIBO	AUC	0.74413 ± 0.0234	0.60070 ± 0.02934	0.62955 ± 0.01989	0.67722 ± 0.02206	0.69628 ± 0.02600
	p-value	-	<0.0001	<0.0001	0.00001	0.00003
CDI	AUC	0.84091 ± 0.01890	0.73271 ± 0.01730	0.68482 ± 0.01535	0.76280 ± 0.01307	0.75288 ± 0.03125
	p-value	-	<0.0001	<0.0001	0.00001	0.00015
LI	AUC	0.81647 ± 0.00751	0.52619 ± 0.02473	0.50884 ± 0.02027	0.77253 ± 0.00730	0.79129 ± 0.00749
	p-value	-	<0.0001	<0.0001	<0.0001	<0.0001
CD	AUC	0.83432 ± 0.01255	0.53990 ± 0.02785	0.57160 ± 0.02203	0.70787 ± 0.02003	0.73538 ± 0.02834
	p-value	-	<0.0001	<0.0001	<0.0001	<0.0001
MD	AUC	0.62984 ± 0.01240	0.57123 ± 0.01472	0.55361 ± 0.01702	0.58390 ± 0.01853	0.58468 ± 0.01997
	p-value	-	0.00005	<0.0001	0.00006	0.0002
UH	AUC	0.70936 ± 0.00442	0.71976 ± 0.00533	0.71829 ± 0.00730	0.75626 ± 0.00859	0.76266 ± 0.00520
	p-value	-	0.00016	0.02248	<0.0001	<0.0001

654 The model training process was repeated 10 times to assess the significance of differences in the
655 model performance. P values were calculated by comparing with AUC scores of models using Meta
656 respectively. The AUC scores were presented as the mean ± standard deviation.

657 *Table 2. Top ten most important features using three types of feature sets for IBD .*

Meta		Meta-OTUab		Meta-OTUoc	
Feature Name	Rank	Feature Name	Rank	Feature Name	Rank
VITAMIN_B_SUPPLEMENT	1.9	VITAMIN_B_SUPPLEMENT	1.2	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;s_	38.3
ALCOHOL	4.7	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coproccoccus;s_	5.1	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_;s_	43.2
VITAMIN_D_SUPPLEMENT	4.7	c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Holdemania;s_	7.8	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_;s_	43.3
BMI	6.2	VITAMIN_D_SUPPLEMENT	12.2	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coproccoccus;s_	45.4
AGE_CORRECTED	8.2	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_;s_	17.8	AGE_CORRECTED	46.2
HEIGHT_CM	9.1	PROBIOTIC	19.5	c_Clostridia;o_Clostridiales;f_;g_;s_	48
WHOLE_GRAIN	9.2	c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g_Haemophilus;s_parainfluenzae	20.3	c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Klebsiella;s_	50.4
PROBIOTIC	10	c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_	25.6	HEIGHT_CM	52.1
WEIGHT_KG	12	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;s_	28.7	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_[Ruminococcus];s_	54.4
RED_MEAT	12.2	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii	31.5	c_Clostridia;o_Clostridiales;f_;g_;s_	56.6

658

659

660 *Table 3. Top 10 features using three types of feature sets for unhealthy status.*

Meta	OTUab	Meta-OTUab	Meta-OTUoc
MILK_CHEESE/1	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii/3.2	MILK_CHEESE/1	MILK_CHEESE/1
PROBIOTIC/2.6	c_Clostridia;o_Clostridiales;f_;g_;s_/4.3	PROBIOTIC/2	PROBIOTIC/2
AGE_CORRECTED/5.9	c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_/10.3	VITAMIN_D_SUPPLEMENT/7.1	c_Clostridia;o_Clostridiales;f_;g_;s_/8.6
MILK_SUBSTITUTE/6.4	c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium;s_adolescentis/10.4	c_Clostridia;o_Clostridiales;f_;g_;s_/7.9	VITAMIN_D_SUPPLEMENT/11
VITAMIN_D_SUPPLEMENT/8.3	c_Clostridia;o_Clostridiales;f_;g_;s_/11.2	MILK_SUBSTITUTE/11.7	MILK_SUBSTITUTE/13.1
LATITUDE/9.2	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;s_/14.1	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii/12.5	FROZEN_DESSERT/18.9
BMI/9.3	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_;s_/14.7	c_Clostridia;o_Clostridiales;f_;g_;s_/14.1	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_;s_/20.8
ELEVATION/10	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii/15.1	AGE_CORRECTED/19.3	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;s_/22.6
FROZEN_DESSERT/10.7	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_;s_/22.4	c_Clostridia;o_Clostridiales;f_;g_;s_/31	AGE_CORRECTED/23.5
WEIGHT_KG/10.7	c_Clostridia;o_Clostridiales;f_;g_;s_/22.6	c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_/39.8	LATITUDE/31.5

661

662

663 *Table 4. Comparing the differences of AUCs of nine diseases using five feature types*
664 *after removing probiotic, vitamin B and D.*

Feature type		Meta	OTUab	OTUoc	Meta-OT Uab	Meta-OT Uoc
IBD	AUC	0.70430 ± 0.01896	0.71616 ± 0.01961	0.70988 ± 0.01362	0.71616 ± 0.01731	0.73755 ± 0.01779
	p-value	-	0.1578	0.57154	0.21862	0.00011
DI	AUC	0.85186 ± 0.01353	0.65737 ± 0.01907	0.64386 ± 0.05136	0.81847 ± 0.01887	0.83323 ± 0.02016
	p-value	-	<0.00001	<0.00001	0.00406	0.07444
IBS	AUC	0.66737 ± 0.00874	0.59746 ± 0.0115	0.57002 ± 0.01385	0.65477 ± 0.00876	0.66202 ± 0.01747
	p-value	-	<0.00001	<0.00001	0.01195	0.51024
SIBO	AUC	0.70714 ± 0.01773	0.60376 ± 0.02748	0.62955 ± 0.01989	0.65295 ± 0.02895	0.67997 ± 0.02833
	p-value	-	<0.00001	<0.00001	0.00002	0.00833
CDI	AUC	0.82321 ± 0.01605	0.72817 ± 0.02593	0.68964 ± 0.01523	0.74306 ± 0.02055	0.73450 ± 0.01714
	p-value	-	<0.00001	<0.00001	0.00001	<0.00001
LI	AUC	0.81425 ± 0.00502	0.52895 ± 0.01525	0.50413 ± 0.02001	0.76971 ± 0.0082	0.78787 ± 0.00734
	p-value	-	<0.00001	<0.00001	<0.00001	0.00001
CD	AUC	0.82741 ± 0.01343	0.54769 ± 0.03572	0.57925 ± 0.01822	0.70966 ± 0.03159	0.73393 ± 0.02467
	p-value	-	<0.00001	<0.00001	<0.00001	<0.00001
MD	AUC	0.63616 ± 0.01439	0.57518 ± 0.01564	0.55468 ± 0.0179	0.57928 ± 0.01538	0.58750 ± 0.02399
	p-value	-	0.00003	<0.00001	0.00016	0.00232
UH	AUC	0.69841 ± 0.00528	0.72057 ± 0.00633	0.71888 ± 0.0072	0.74820 ± 0.00376	0.75979 ± 0.00446
	p-value	-	0.00007	0.00008	<0.00001	<0.00001

665 P values were calculated by comparing models' AUC scores with the models using Meta respectively
666 and the AUC score was denoted as the mean value ± standard deviation.

668 **Supplementary Tables**

669 *Table S1. Frequency of diet and lifestyle factors.*

	Never	Rarely	Occasionally	Regularly	Daily
ALCOHOL	1643 (21.70%)	1889 (24.95%)	1794 (23.69%)	1459 (19.27%)	786 (10.38%)
POULTRY	840 (11.09%)	934 (12.33%)	3021 (39.90%)	2531 (33.43%)	245 (3.236%)
EXERCISE	211 (2.786%)	812 (10.72%)	1934 (25.54%)	3002 (39.65%)	1612 (21.29%)
FERMENTED_P LANT	2501 (33.03%)	2576 (34.02%)	1136 (15.00%)	804 (10.61%)	554 (7.317%)
FROZEN_DESS ERT	2132 (28.16%)	3764 (49.71%)	1204 (15.90%)	418 (5.521%)	53 (0.700%)
FRUIT	376 (4.966%)	997 (13.16%)	1542 (20.36%)	2333 (30.81%)	2323 (30.68%)
HIGH_FAT_RED _MEAT	1608 (21.23%)	3032 (40.04%)	2007 (26.50%)	833 (11.00%)	91 (1.201%)
HOMECOOKED _MEALS	123 (1.624%)	199 (2.628%)	466 (6.155%)	2562 (33.83%)	4221 (55.75%)
RED_MEAT	1135 (14.99%)	1823 (24.07%)	2677 (35.35%)	1777 (23.47%)	159 (2.100%)
MEAT_EGGS	300 (3.962%)	388 (5.124%)	1184 (15.63%)	3067 (40.50%)	2632 (34.76%)
MILK_SUBSTIT UTE	3578 (47.25%)	1250 (16.51%)	824 (10.88%)	792 (10.46%)	1127 (14.88%)
MILK_CHEESE	1116 (14.74%)	1328 (17.54%)	1424 (18.80%)	1986 (26.23%)	1717 (22.67%)
OLIVE_OIL	468 (6.181%)	720 (9.509%)	1557 (20.56%)	2870 (37.90%)	1956 (25.83%)
PROBIOTIC	3148 (41.57%)	1501 (19.82%)	728 (9.615%)	762 (10.06%)	1432 (18.91%)
SALTED_SNAC KS_FREQUENC Y	1242 (16.40%)	2718 (35.90%)	2086 (27.55%)	1202 (15.87%)	323 (4.266%)
SEAFOOD	911 (12.03%)	2253 (29.75%)	3329 (43.97%)	1012 (13.36%)	66 (0.871%)
SMOKING	7109 (93.89%)	264 (3.486%)	58 (0.766%)	36 (0.475%)	104 (1.373%)
VEGETABLE	51 (0.673%)	204 (2.694%)	674 (8.902%)	2661 (35.14%)	3981 (52.58%)
VITAMIN_D_SU PPLEMENT	3320 (43.85%)	933 (12.32%)	536 (7.079%)	776 (10.24%)	2006 (26.49%)
VITAMIN_B_SU PPLEMENT	4301 (56.80%)	909 (12.00%)	472 (6.234%)	542 (7.158%)	1347 (17.79%)
WHOLE_GRAIN	945 (12.48%)	1520 (20.07%)	1669 (22.04%)	2102 (27.76%)	1335 (17.63%)
WHOLE_EGGS	603 (7.964%)	1208 (15.95%)	2913 (38.47%)	2013 (26.58%)	834 (11.01%)

670 Rarely: less than once/week; Occasionally: 1–2 times/week; Regularly: 3–5 times/week.

671

672 *Table S2. Basic information of the dataset.*

	Female	Male	Total
Number	3983	3588	7571
Age	47(35,59)	46(33,61)	47(34,59)
BMI	22.43(20.36,25.63)	24.28(22.05,26.61)	23.4(20.98, 26.3)
Height (cm)	165(160,170)	178(172,183)	170(162,178)
Weight (kg)	61(55,70)	78(69,87)	69(58,81)
Latitude	40.8(34.1,51.3)	40.1(33,50.175)	40.7(33.6,51)
Elevation	81.2(28.4,193.3)	96.1(30.9,205.375)	88.1(29.8,200.6)
Caucasian	3575	3186	6761
Asian or Pacific Islander	162	178	340
African American	29	16	45
Hispanic	97	102	199
Other Race	120	106	226
CARD	74	180	254
SIBO	209	101	310
MENTAL	407	270	677
LACTOSE	712	508	1220
IBS	666	333	999
IBD	164	249	413
Diabetes	74	100	174
CDIFF	92	87	179
Unhealth	1612	1309	2921

673 The representation format of the values for age, BMI, height, weight, latitude and elevation are median
674 (lower quartile and upper quartile).

675

676 *Table S3. Top 10 features using three types of feature sets for Diabetes.*

Meta		Meta-OTUab		Meta-OTUoc	
Feature Name	Rank	Feature Name	Rank	Feature Name	Rank
BMI	1.2	BMI	2.5	BMI	1.6
AGE_CORRECTED	2	AGE_CORRECTED	16.4	AGE_CORRECTED	9.2
ELEVATION	6.1	c_Deltaproteobacteria;o_Desulfovibrionales;f_Desulfovibrionaceae;g_Bilophila;s_	28.8	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;s_	18.7
LATITUDE	8.2	c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces;s_	30	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_;;s_	19.4
WEIGHT_KG	8.8	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_;;s_	31.1	MILK_CHEESE_FREQUENCY	27.5
ALCOHOL_FREQUENCY	8.9	c_Bacteroidia;o_Bacteroidales;f_[Paraprevotellaceae];g_Paraprevotella;s_	44.1	ALCOHOL_FREQUENCY	35.1
MILK_CHEESE_FREQUENCY	10	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_;;s_	48	HEIGHT_CM	41.1
EXERCISE_FREQUENCY	11.5	c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Streptococcus;s_	48.8	LATITUDE	41.5
HEIGHT_CM	11.5	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_;;s_	50.2	ELEVATION	41.6
MILK_SUBSTITUTE_FREQUENCY	11.8	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;s_	51.1	MEAT_EGGS_FREQUENCY	42.1

677

678

679 *Table S4. Top 10 features using two types of feature sets for IBS.*

Meta		Meta-OTUoc	
Feature Name	Rank	Feature Name	Rank
LATITUDE	1.8	MILK_CHEESE	3.8
AGE_CORRECTED	2.5	PROBIOTIC	4
PROBIOTIC	4.6	LATITUDE	4.9
MILK_CHEESE	5	AGE_CORRECTED	8.7
WEIGHT_KG	6.6	WEIGHT_KG	10.7
ELEVATION	6.8	c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Enterobacter;s_radicincitans	17
BMI	7.1	female	26.9
EXERCISE	10.7	HEIGHT_CM	28
HEIGHT_CM	11	WHOLE_GRAIN	28.8
SEAFOOD	11.5	SEAFOOD	30.3

680

681

682 *Table S5. Top 10 features using human variables for the other five diseases.*

SIBO		CDI		LI		CD		MD	
Feature Name	Rank	Feature Name	Rank	Feature Name	Rank	Feature Name	Rank	Feature Name	Rank
PROBIOTIC_FREQUENCY	1.2	WHOLE_GRAIN_FREQUENCY	2.3	MILK_CHEESE_FREQUENCY	1.2	AGE_CORRECTED	1.4	AGE_CORRECTED	2
HEIGHT_CM	2.8	MILK_CHEESE_FREQUENCY	3.6	MILK_SUBSTITUTE_FREQUENCY	1.8	female	6	BMI	4.9
LATITUDE	7.2	WEIGHT_KG	4.3	Caucasian	5.1	HEIGHT_CM	6.5	HEIGHT_CM	6.2
BMI	7.8	PROBIOTIC_FREQUENCY	6.2	ALCOHOL_FREQUENCY	6	RED_MEAT_FREQUENCY	6.7	WEIGHT_KG	7.8
VITAMIN_B_SUPPLEMENT_FREQUENCY	8.4	BMI	8.6	POULTRY_FREQUENCY	8.4	WEIGHT_KG	7	MILK_SUBSTITUTE_FREQUENCY	8.1
ELEVATION	9.1	LATITUDE	8.8	HEIGHT_CM	9.4	OLIVE_OIL	9.2	ALCOHOL_FREQUENCY	8.2
AGE_CORRECTED	10.2	HEIGHT_CM	9.6	HIGH_FAT_RED_MEAT_FREQUENCY	9.9	LATITUDE	9.6	LATITUDE	8.4
ALCOHOL	10.5	FROZEN	10.5	SALT	10.6	BMI	10.2	ELEVATION	10.8

_FREQUENCY		_DESSERT_FREQUENCY		ED_SNACKS_FREQUENCY				ON	
WHOLE_EGGS	11.5	POULTRY_FREQUENCY	11.2	PROBIOTIC_FREQUENCY	12.4	ELEVATION	11.8	VITAMIN_D_SUPPLEMENT_FREQUENCY	12.6
MILK_CHEESE_FREQUENCY	13.1	AGE_CORRECTED	11.3	ELEVATION	12.5	FRUIT_FREQUENCY	11.9	FERMENTED_PLANT_FREQUENCY	13.1

684 *Table S6. Top 10 features after removing probiotics, vitamin B and vitamin D for all*
685 *diseases.*

IBD	DI	IBS	SIBO	CDI	LI	CD	MD	UH
Meta-O TUoc	Meta-OT Uoc	Meta-OT Uab	Meta-OT Uab	Meta	Meta		Meta- OTUo c	Meta- OTUo c
BMI/45 .3	BMI/1.7	MILK_CH EESE_FR EQUENC Y/3.4	MILK_CH EESE_FR EQUENC Y/10.5	HEIGHT_C M/1.6	MILK _CHE ESE_ FRE QUEN CY/1. 1	AGE_ CORR ECTE D/1.5	MILK _SUB STITU TE_FR EQUE NCY/5 .4	MILK _CHE ESE/1
HEIGH T_CM/ 48.3	AGE_CO RRECTE D/7.2	female/4.9	WHOLE_ GRAIN_F REQUEN CY/16.8	BMI/5.3	MILK _SUB STIT UTE_ FRE QUEN CY/1. 9	female /2.2	AGE_ CORR ECTE D/28.2	c_Clos tridia;o _Clostr idiales; f_;g_;s _/7.8
c_Clostr idia;o_ Clostrid iales;f_ Rumino coccace ae;g_Ru minoco ccus;s_ /48.4	c_Clostrid ia;o_Clost ridiales;f_ Ruminoco ccaceae;g _;s_/17.8	c_Gammap roteobacter ia;o_Enter obacteriale s;f_Enterobacteriaceae;g_Enterobacter;s_radicincitans/ 5.3	FROZEN_ DESSERT _FREQUE NCY/17	MILK_CH EESE_FRE QUENCY/6 .3	ELEV ATIO N/7.3	RED_ MEAT _FRE QUEN CY/5.2	ELEV ATIO N/34.4	MILK _SUB STITU TE/10. 3
c_Clostr idia;o_ Clostrid iales;f_ Rumino coccace ae;g_Ru minoco ccus;s_ /48.8	WEIGHT _KG/22.6	AGE_COR RECTED/ 9.6	c_Gammap roteobacter ia;o_Enter obacteriale s;f_Enterobacteriaceae;g_Enterobacter;s_radicincitans/ 28.7	FROZEN_ DESSERT_ FREQUEN CY/6.7	ALC OHO L_FR EQU ENC Y/7.7	HEIG HT_C M/6.4	HEIG HT_C M/35.5	c_Clos tridia;o _Clostr idiales; f_;g_;s _/15.2
c_Clostr idia;o_ Clostrid iales;f_ Lachnos	c_Clostrid ia;o_Clost ridiales;f_ Ruminoco ccaceae;g	LATITUD E/10.5	c_Clostridi a;o_Clostri diales;f_Ru minococca ceae;g_Osc	AGE_COR RECTED/9. 5	HEIG HT_C M/8.5	WEIG HT_K G/7.5	BMI/3 5.9	c_Clos tridia;o _Clostr idiales; f_Rum

piraceae; g;s_/50.3	_Ruminococcus;s_/23.3		illospira;s_/32.1					inococ caceae; g_Rum inococ cus;s_/15.2
c_Clostridia;o_Clostridiales;f_[Tissierellaceae]; g_Anaerococcus;s_/50.4	ALCOHOL_FREQ UENCY/26.7	SEAFOOD_FREQ NCY/25.4	c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Veillonella;s_parvula/44.4	LATITUDE/10.9	BMI/8.5	FRUIT_FREQ QUENCY/8.6	ALCOHOL_FREQ UENCY/42.3	FROZEN_D ESSE RT/19.8
c_Clostridia;o_Clostridiales;f_g;s_/54.1	MILK_CHEESE_FREQ REQUENCY/34.2	MILK_SUBSTITUTE_FREQ ENCY/27.8	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g;s_/49.7	FERMENTED_PLAN T_FREQ ENCY/11.4	LATITUDE/8.6	OLIVE_OIL/11.6	WEIGHT_KG/46.7	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g;s_/21.8
c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g;s_/54.2	ELEVATION/34.7	WEIGHT_KG/32.4	POULTRY_FREQ NCY/55.1	ALCOHOL_FREQ NCY/11.5	Caucasian/9	LATITUDE/1.6	FROZEN_D ESSE RT_FREQ REQUENCY/48.9	AGE_CORRECTED/22.6
c_Clostridia;o_Clostridiales;f_g;s_/60.6	HEIGHT_CM/36.1	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g;s_/40.2	c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_[Ruminococcus];s_/55.8	MILK_SUBSTITUTE_FREQ CY/13	AGE_CORRECTED/9.7	BMI/1.6	c_Clostridia;o_Clostridiales;f_Clostridiaceae;g;s_/60.2	BMI/23.9
MILK_CHEESE/66.3	c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Clostr	HEIGHT_CM/49.9	c_Clostridia;o_Clostridiales;f_g;s_/56.6	WHOLE_EGGS/13.3	SALTED_SNAKS_FREQ UENCY/1	SEAFOOD_FREQ UENCY/13.2	c_Clostridia;o_Clostridiales;f_Clostridiaceae;g;s	c_Clostridia;o_Clostridiales;f_Ruminococcaceae;

	idium;s_s accharogu mia/36.9				0.1		_/63.4	g_;s_/2 4.9
--	------------------------------------	--	--	--	-----	--	--------	----------------

A

- ①: Using human lifestyle and physiological variables to classify human disease with machine learning methods;
- ②: Using human variables and gut microbiota to classify human disease with machine learning methods;
- Only when ② > ①, independent association of gut microbiota with disease can be confirmed;
- Association strengths between gut microbiota and disease is evaluated by classification performance of ML models

B