The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features for human complex diseases

Fengzhe Xu[1#], Yuanqing Fu[1#], Ting-yu Sun[2], Zengliang Jiang[1,3], Zelei Miao[1], Menglei Shuai[1], Wanglong Gou[1], Chu-wen Ling[2], Jian Yang[4,5], Jun Wang[6]*, Yu-ming Chen[2]*, Ju-Sheng Zheng[1,3,7]*

[#]These authors contributed equally to the work

[1] School of Life Sciences, Westlake University, Hangzhou, China.

[2] Guangdong Provincial Key Laboratory of Food, Nutrition and Health; Department of Epidemiology, School of Public Health, Sun Yat-sen University, Guangzhou, China.

[3] Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, China.

[4] Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia.

[5] Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China

[6] CAS Key Laboratory for Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China.

[7] MRC Epidemiology Unit, University of Cambridge, Cambridge, UK.

Short title: Interplay between and host genetics and gut microbiome

*Correspondence to

Prof Ju-Sheng Zheng

School of Life Sciences, Westlake University,18 Shilongshan Rd, Cloud Town, Hangzhou, China. Tel: +86 (0)57186915303. Email: zhengjusheng@westlake.edu.cn

And

Prof Yu-Ming Chen

Guangdong Provincial Key Laboratory of Food, Nutrition and Health; Department of Epidemiology, School of Public Health, Sun Yat-sen University, Guangzhou, China. Email: chenyum@mail.sysu.edu.cn

And

Prof Jun Wang

CAS Key Laboratory for Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. Email: junwang@im.ac.cn

Abstract

There is increasing interest about the interplay between host genetics and gut microbiome on human complex diseases, with prior evidence mainly derived from animal models. In addition, the shared and distinct microbiome features among human complex diseases remain largely unclear. We performed a microbiome genome-wide association study to identify host genetic variants associated with gut microbiome in a Chinese population with 1475 participants. We then conducted bi-directional Mendelian randomization analyses to examine the potential causal associations between gut microbiome and human complex diseases. We did not find evidence supporting the causal effect of gut microbiome on human complex diseases. In contrast, atrial fibrillation, chronic kidney disease and prostate cancer, as predicted by the host genetics, had potential causal effect on gut microbiome. Further disease-microbiome feature analysis suggested that gut microbiome features revealed novel relationship among human complex diseases. These results suggest that different human complex diseases share common and distinct gut microbiome features, which may help re-shape our understanding about the disease etiology in humans.

Introduction

Ever-increasing evidence has suggested that gut microbiome is involved in many physiological processes, such as energy harvest, immune response, and neurological function[1-3]. With successes of investigation into the clinical application of fecal transplants, modulation of gut microbiome has emerged as a potential treatment option for some complex diseases, including inflammatory bowel disease and colorectal cancer [4,5]. However, it is still unclear whether the gut microbiome has the potential to be clinically applied for the prevention or treatment of many other complex diseases. Therefore, it is important to clarify the bi-directional causal association between gut microbiome and human complex diseases or traits.

Mendelian randomization (MR) is a method that uses genetic variants as instrumental variables to investigate the causality between an exposure and outcome in observational studies[6]. Prior literature provides evidence that the composition or structure of the gut microbiome can be influenced by the host genetics[7-10]. On the other hand, host genetic variants associated with gut microbiome were rarely explored in Asian populations, thus we are still lacking instrument variables to perform MR for gut microbiome in Asians. This calls for novel microbiome genome-wide association study (GWAS) in Asian populations.

Along with the causality issue between the gut microbiome and human complex diseases, it is so far unclear whether human complex diseases had similar or unique

gut microbiome features. Identifying common and distinct gut microbiome features across different diseases might shed light on novel relationships among the complex diseases and update our understanding about the disease etiology in humans. However, the composition and structure of gut microbiome are influenced by a variety of factors including environment, diet and regional variation [11-13], which posed a key challenge for the description of representative microbiome features for a specific disease. Although there were several studies comparing disease-related gut microbiome [14-16], few of them has examined and compared the microbiome features across different human complex diseases.

In the present study, we performed a microbiome GWAS in a Chinese cohort study: the Guangzhou Nutrition and Health Study (GNHS)[17], including 1475 participants. Subsequently, we applied a bi-directional MR method to explore the genetically predicted relationship between gut microbiome and human complex diseases. To explore novel relationships among human complex diseases based on gut microbiome, we investigated the shared and distinct gut microbiome features across diverse human complex diseases [18].

Result

Overview of the study

Our study was based on the GNHS, with 4048 participants (40-75 years old) living in urban Guangzhou city recruited during 2008 and 2013 [17]. In the GNHS, stool samples

were collected among 1937 participants during follow-up visits, among which 1475 unrelated participants without taking anti-biotics were included in our discovery microbiome GWAS. We then included additional 199 participants with both genetic and gut microbiome data as a replication cohort, which belonged to the control arm of a case-control study of hip fracture in Guangdong Province, China [19].

For both discovery and replication cohorts, genotyping was carried out with Illumina ASA-750K arrays. Quality control and relatedness filters were performed by PLINK1.9 [20]. We conducted the genome-wide genotype imputation with 1000 Genomes Phase3 v5 reference panel by Minimac3[21-23]. HLA region was imputed with Pan-Asian reference panel and SNP2HLA v1.0.3 [24-26].

Association of host genetics with gut microbiome features

We performed a series of microbiome GWAS with PLINK 1.9 based on logistic models for binary variables [20]. For continuous variables, we used GCTA with mixed linear model-based association (MLMA) method [27,28]. We also analyzed categorical variable enterotypes of the participants based on genus-level relative abundance of gut microbiome, using the Jensen-Shannon Distance (JSD) and the Partitioning Around Medoids (PAM) clustering algorithm [29]. The participants were subsequently clustered into two groups according to the enterotypes (*Prevotella* vs *Bacteroides*). Thereafter, we performed GWAS for enterotypes using logistic regression model to explore potential associations between host genetics and enterotypes. However, we did not

find any genome-wide significant locus with $p<5\times10^{-8}$. Furthermore, we used a restricted maximum likelihood analysis (REML) with GCTA to estimate the SNP-based heritability, and the estimate heritability of the enterotype was 0.055 (SE=0.19, Supplementary Table S2) [30].

To examine the association of host genetic variants with alpha diversity, we performed GWAS for three indices (Shannon diversity index, Chao1 diversity indices and observed OTUs index), but again no genome-wide significant signal ($p<5\times10^{-8}$) was found. In the discovery cohort, the heritability of alpha diversity ranged from 0.054 to 0.14 (SE=0.20 for all indices, Supplementary Table S2). To further investigate if there is host genetic basis underlying alpha diversity, we constructed a polygenic score for each alpha diversity indicator in the replication cohort, using the genetic variants which showed suggestive significance ($p<5\times10^{-5}$) in the discovery GWAS. The polygenic score was not significantly associated with its corresponding alpha diversity index in our replication cohort. Meanwhile, none of the associations with alpha diversity indices reported in the literature could be replicated (Supplementary Table S7) [7].

We performed a beta diversity GWAS using a tool called MicrobiomeGWAS [31], and found that one locus at *SMARCA2* gene (rs6475456) was associated with beta-diversity at a genome-wide significance level ($p=3.96\times10^{-9}$). However, we could not replicate the results in the replication cohort, which may be due to the limited

sample size of the replication cohort. In addition, prior literature had reported 73 genetic variants that were associated with beta diversity [8,13,32,33], among which we found that 3 single nucleotide polymorphisms (SNP, *UHRF2* gene-rs563779, *LHFPL3* gene-rs12705241, *CTD-2135J3.4*-rs11986935) had nominal significant ($p<0.05$) association with beta-diversity in our cohort (Supplementary Table S6), although none of the association survived Bonferroni correction. These studies used various methods for the sequencing and calculation of beta diversity, which raised challenges to verify and extrapolate results across populations.

We then took the genetic loci reported to be associated with individual taxa in prior studies [7,8,13,33] for replication in our GNHS dataset. Although there are still some signals with nominal significance ($p<0.05$) in our study (e.g., 7 loci associated with *Lachnospiraceae*, *Coprococcus* or *Bacteroides* with $p<0.05$; Supplementary Table S5), none of the associations of these genetic variants with taxa survived the Bonferroni correction ($p<1\times10^{-4}$). The null results may be because of various clustering similarities, classifiers or reference databases to annotate taxa and different sequencing methods used in these studies.

We subsequently performed GWAS discovery for individual gut microbes in our own GNHS discovery dataset. For the taxa present at more than ninety percent of participants and alpha diversity, we used Z-score normalization to transform the distribution and carried out analysis based on a log-normal model. A MLMA test in

the GCTA was used to assess the association, with the first five principal components, age, sex and sequencing batch fitted as fixed effects and the effects of all the SNPs fitted as random effects [27,28,34]. For other taxa present at fewer than ninety percent, we transformed the absence/presence of the taxon into binary variables and used PLINK1.9 to perform a logistic model, adjusted for the first five principal components, age, sex and sequencing batch.

For all the gut microbiome taxa, the significant threshold was defined as 5×10^{-8} in the discovery stage. As some taxa were correlated with each other, we also used an eigendecomposition analysis to calculate the effective number of independent taxa on each taxonomy level (phylum level: 2.3, class level: 2.9, order level: 2.9, family level: 5.5, genus level: 5.6, species level: 3.2) [35,36]. We found that 6 taxa were associated with host genetic variants in the discovery cohort ($p<5\times10^{-8}/n$, n is the effective number of independent taxa on each taxonomy level, Supplementary Table S4); however, these associations were not significant ($p>0.05$) in the replication cohort. We then used a threshold of $p<5\times10^{-5}$ at the GWAS discovery stage to incorporate additional genetic variants which may explain a larger proportion of heritability for taxa, based on which we constructed a polygenic score for each taxon in the replication. We found that the polygenic scores were significantly associated with 3 taxa including *Coriobacteriaceae*, *Odoribacter* and *Parabacteroides_undefined* in the replication set ($p<0.05$, Methods, see also Figure 1A, 1B, 1C).

Genetic correlation of gut microbiome and traits

We used GCTA to perform a bivariate GREML (genomic-relatedness-based restricted maximum-likelihood) analysis to estimate the genetic correlation between gut microbiome and traits in the GNHS [27,37]. The traits included BMI, fasting blood sugar (FBS), glycosylated hemoglobin (HbA1c), systolic blood pressure (SBP), diastolic blood pressure (DBP), high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglyceride (TG), none of which could pass Bonferroni correction. Additionally, HDL-C was the only trait that had nominal genetic correlation ($p<0.05$) with gut microbes (specifically, *Desulfovibrionaceae* and *[Prevotella]*, Supplementary Table S3).

Bi-directional assessment of the genetically predicted association between gut microbiome and complex diseases/traits

Using genetic variants-composed polygenic scores as genetic instruments, we performed MR analysis to assess the putative causal effect of microbiome (*Coriobacteriaceae*, *Odoribacter* and *Parabacteroides_undefined*) on human complex diseases or traits. Inverse variance weighted (IVW) method was used for the MR analysis, while other three methods (Weighted median, MR-Egger and MR-PRESSO) [38,39] were applied to confirm the robustness of results. The horizontal pleiotropy was assessed using MR-PRESSO Global test and MR-Egger Regression. For the analysis of gut microbiome on complex traits, we downloaded public available GWAS summary statistics of complex traits (n=58) and diseases (type 2 diabetes mellitus

(T2DM), atrial fibrillation (AF), colorectal cancer (CRC) and prostatic cancer (PCa)) reported by BioBank Japan [40-44]. There was no evidence that these taxa had causal association with human complex diseases or traits in our MR analyses (Supplementary Table S9), which may be due to the limited genetic instruments discovered in our present study.

We subsequently performed a reserve MR analysis to assess the potential causal effect of human complex diseases on gut microbiome features. For the reserve MR analyses, the diseases of interests included T2DM, AF, coronary artery disease (CAD), chronic kidney disease (CKD), Alzheimer's disease (AD), CRC and PCa, and their instrumental variables for the MR analysis were based on the previous large-scale GWAS in East Asians[40,45-50]. The results suggested that AF and CKD were causally associated with gut microbiome (See also Figure 2A, 2B, Supplementary Table S10). Genetically predicted higher risk of AF was associated with lower abundance of *Coprophilus*, *Lachnobacterium*, *Barnesiellaceae*, *Veillonellaceae* and *Mitsuokella*, and higher abundance of *Alcaligenaceae*. Additionally, genetically predicted higher risk of CKD could increase *Anaerostipes* abundance and higher risk of PCa could decrease *[Prevotella]*.

To further investigate the potential complex diseases that may be correlated with the taxa affected by AF, we applied Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) to predict the disease pathway

abundance [51]. We used Spearman's rank-order correlation to test whether 22 human complex diseases were associated with the aforementioned AF-associated taxa (See also Figure 2C). The heatmap indicated that cancers and neurodegenerative diseases including Parkinson's disease (PD), AD, amyotrophic lateral sclerosis (ALS) as well as AF were correlated with similar gut microbiome. Although the association among these diseases are highly supported by previous studies [52-54], no study has compared common gut microbiome features across these different diseases.

Microbiome features of human complex diseases

To compare gut microbiome features across human diseases, we chose 22 human complex diseases from predicted abundance and performed k-medoids clustering [18]. We used an m × n matrix to perform the cluster analysis, where m is the number of participants and n is the number of selected diseases. According to optimum average silhouette width [55], we chose optimal number of clusters for further analysis (See also Figure 3A). The plot showed that neurological diseases including ALS and AD belonged to the same cluster, while PD and CRC had much similarity in gut microbiome. The results also suggested that systemic lupus erythematosus (SLE) and chronic myeloid leukemia (CML) shared similar gut microbiome features. Moreover, we could replicate these clusters in our replication cohort, which suggested that the clustering results were robust (See also Figure 3B).

We further asked whether gut microbiome contributed to the novel clustering. To this

end, we repeated the analysis among participants who took antibiotic less than two weeks before stool sample collection, considering that antibiotic treatments were believed to cause microbiome imbalance, and the clusters were totally different in this group (See also Figure 3C). The results indicated a totally different clustering, which suggested that gut microbiome indeed contributed to the correlations among diseases. To further demonstrate common microbiome features across different diseases, we examined the correlation of the predicted diseases with genus-level taxa. The results showed that human complex diseases had shared similar gut microbiome features, as well as distinct features on their own (See also Figure 4).

To validate the accuracy of the association between the predicted disease-related gut microbiome features and the corresponding disease, we used T2DM as an example, examining the association of predicted T2DM-related microbiome features with T2DM risk in our GNHS samples. We constructed a microbiome risk score (MRS) based on 16 selected taxa with predicted correlation coefficient with T2DM greater than 0.2. A logistic regression was used to examine the above MRS with T2DM risk in the GNHS (n=1886, with 217 T2DM cases). The results showed that higher MRS was associated with lower risk of T2DM (odds ratio: 0.850, 95% confidence interval: 0.804 to 0.898, $p=8.75\times10^{-9}$).

Based on the above results, we proposed a hypothesis that related diseases might share similar gut microbiome features. To test for this hypothesis, we performed

validation analysis by including GNHS participants who had one of the following self-reported diseases: stroke (n=8), chronic hepatitis (n=19), coronary heart diseases (CHD) (n=40), cataract (n=124) and insomnia (n=68). The results of k-medoids clustering suggested that CHD, cataract and insomnia shared common gut microbiome features, which was supported by the prior research reporting that both patients suffering insomnia and women receiving cataract extraction had increased risks of CHD [56-58].

Discussion

Our study is among the first to investigate the host genetics-gut microbiome association in East Asian populations and reveals that several microbiome species (e.g., *Coriobacteriaceae* and *Odoribacter*) are influenced by host genetics. We then show that complex diseases such as atrial fibrillation, chronic kidney disease and prostate cancer, have potential causal effect on gut microbiome. More interestingly, our results indicate that different human complex diseases may be mechanically correlated by sharing common gut microbiome features, but also maintaining their own distinct microbiome features.

Previous studies and our study showed that gut microbiome had an inclination to be influenced by host genetics [8,10,33,59], although the successful replication tends to be rare. We could not validate any of the reported genetic variants that were significantly associated with gut microbiome features in prior reports, which may reflect the

298 difference in population and heterogeneity between study but also raise concerns
299 about the reproducibility. Many factors including ethnic differences,
300 gene-environment interaction and dissimilarity in sequencing methods may make it
301 hard to extrapolate results from microbiome GWAS across populations in the
302 microbiome field. Nevertheless, we successfully replicate several polygenic scores of
303 gut microbiome, and the current study represent the largest dataset in Asian
304 populations and would be a unique resource to be used in large-scale trans-ethnic
305 meta-analysis of microbiome GWAS in future.
306
307 The MR analysis in the present study did not support causal effect of gut microbiome
308 on diseases or traits, however, this result should be interpreted with caution because of
309 the limited genetic instruments derived from GWAS. In contrast, the reverse MR
310 analysis provided evidence that AF, CKD and PCa could causally influence gut
311 microbiome. As our study is among the first to investigate gene-microbiome
312 association in East Asians, we need further study in this region to confirm our results.
313 Additionally, rare and low-frequency variants may have an important impact on
314 common diseases[60], thus it will be of interest to clarify the effects of low-frequency
315 variants on gut microbiome in cohorts with large sample sizes in future.
316
317 Our results indicate that gut microbiome helps reveal novel and interesting
318 relationships among human complex diseases, suggesting that different diseases may
319 have common and distinct gut microbiome features. A prior study including

participants from different countries identified three microbiome clusters[29]. Notably, this study focused on classifying the individuals into distinct enterotypes regardless of the individuals' health status, while in the present study we described representative microbiome features for diseases of interest. The microbiome features revealed a close association of AF with neurodegenerative diseases as well as cancers, which was supported by prior studies showing that AF had correlation with AD and PD[52,53], and AF patients had relatively higher risks of several cancers including lung cancer and CRC[54,61]. We also observed that microbiome features of SLE and CML were highly similar. Interestingly, a tyrosine kinase inhibitor of platelet-derived growth factor receptor, imatinib, was widely used to treat CML and significantly ameliorated survival in murine lupus autoimmune disease[62]. In addition, association between CRC and PD has been reported in several observational cohorts[63,64]. Collectively, these findings strongly support our hypothesis that human complex diseases sharing similar microbiome features might be mechanically correlated. Furthermore, from the perspectives of risk genes of AF and neurodegenerative diseases, previous GWAS for AF have identified two loci at *PITX2* gene-rs6843082 and *C9orf3* gene-rs7026071, which were also associated with the risk of ALS (p=0.0138 and p=0.049, respectively) [65-67].

In summary, we perform bi-directional MR analyses to examine the causal relationship between gut microbiome and human complex diseases, revealing that some complex diseases causally affect abundance of specific gut microbes. There is

342 no convincing evidence supporting the causal role of gut microbiome on human
343 complex diseases. The disease and gut microbiome association analysis reveals novel
344 relationships among human complex diseases, which may help re-shape our
345 understanding about the disease etiology, as well as extending clinical indications of
346 existing drugs for different diseases.
347
348 **Method**
349 **Study participants and sample collection**
350 Our study was based on the Guangzhou Nutrition and Health Study (GNHS), with
351 4048 participants (40-75 years old) living in urban Guangzhou city recruited during
352 2008 and 2013 [17]. We followed up participants every three years. In the GNHS, stool
353 samples were collected among 1937 participants during follow-up visits. Among
354 those with stool samples, 1717 participants had genetic data and 1475 participants
355 with identical by decent (IBD) less than 0.185.
356
357 We included 199 participants with both genetic and gut microbiome data as a
358 replication cohort, which belonged to the control arm of a case-control study of hip
359 fracture with the participants (52-83 years old) recruited between June 2009 and
360 August 2015 in Guangdong Province, China [19].
361
362 Blood samples of all participants were collected after an overnight fasting and buffy
363 coat was separated from whole blood and stored at -80℃. Stool samples were

collected during the on-site visit of the participants at Sun Yat-sen University. All samples were manually stirred, separated into tubes and stored at -80℃ within four hours.

Genotyping data

For both discovery and replicattion cohorts, DNA was extracted from leukocyte using the TIANamp® Blood DNA Kit as per the manufacturer's instruction. DNA concentrations were determined using the Qubit quantification system (Thermo Scientific, Wilmington, DE, US). Extracted DNA was stored at -80°C. Genotyping was carried out with Illumina ASA-750K arrays. Quality control and relatedness filters were performed by PLINK1.9 [20]. Individuals with high or low proportion of heterozygous genotypes (outliers defined as 3 standard deviation) were excluded[68]. Individuals who had different ancestries (the first two principal components ±5 standard deviation from the mean) or related individuals (IBD>0.185) were excluded[68]. Variants were mapping to the 1000 Genomes Phase3 v5 by SHAPEIT [23,69] and then we conducted the genome-wide genotype imputation with 1000 Genomes Phase3 v5 reference panel by Minimac3[21,22]. Genetic variants with imputation accuracy RSQR > 0.3 and MAF > 0.05 were included in our analysis. We used Pan-Asian reference panel consist of 502 participants and SNP2HLA v1.0.3 to impute HLA region [24-26].

Sequencing and processing of 16S rRNA data

Microbial DNA was extracted from fecal samples using the QIAamp® DNA Stool Mini Kit per the manufacturer's instruction. DNA concentrations were determined using the Qubit quantification system. The V3-V4 region of the 16S rRNA gene was amplified from genomic DNA using primers 341F and 805R. Sequencing was performed using MiSeq Reagent Kits v2 on the Illuimina MiSeq System.

Fastq-files were demultiplexed by the MiSeq Controller Software. Ultra-fast sequence analysis (USEARCH) was performed to trim the sequence for amplification primers, diversity spacers, sequencing adapters, merge-paired and quality filter[70]. Operational taxonomic units (OTUs) were clustered based on 97% similarity using UPARSE[71]. OTUs were annotated with Greengenes 13_8 (https://greengenes.secondgenome.com/). After randomly selecting 10000 reads for each sample, Quantitative Insights into Microbial Ecology (QIIME) software version 1.9.0 was used to calculate alpha diversity (Shannon diversity index, Chao1 diversity indices and observed OTUs index) based on the rarefied OTU counts[72].

Statistical analysis

Genome-wide association analysis of gut microbiome features

For each of the GNHS participants and the replication cohort, we clustered participants based on genus-level relative abundance, estimating the JSD distance and PAM clustering algorithm, and then defined two enterotypes according to Calinski-Harabasz Index[29,73]. PLINK 1.9 was used to perform a logistic regression

model for enterotypes and taxa present at fewer than ninety percent, adjusted for age, sex and the first five principal components.

For beta diversity, the analyses for the genome-wide host genetic variants with beta diversity was performed using MicrobiomeGWAS[31], adjusted for covariates including the first five principal components, age and sex. We filtered OTUs present at fewer than ten percent of participants to calculate Bray–Curtis dissimilarity.

Alpha diversity was calculated after randomly sampling 10000 reads per sample. For the taxa present at more than ninety percent of participants and alpha diversity, we used Z-score normalization to transform the distribution and carried out analysis based on a log-normal model. A mixed linear model based association (MLMA) test in GCTA was used to assess the association, fitting the first five principal components, age, sex and sequencing batch as fixed effects and the effects of all the SNPs as random effects [27,28,34]. For other taxa present at fewer than ninety percent, we transformed the absence/presence of the taxon into binary variables and used PLINK1.9 to perform a logistic model, adjusted for the first five principal components, age, sex and sequencing batch. For all the gut microbiome features, the significant threshold was defined as 5×10^{-8} /n (n is the effective number of independent taxa on each taxonomy level) in the discovery stage. We estimated genomic inflation factors with LDSC v1.0.1 at local server [74].

Proportion of variance explained by all SNPs

We used the GREML method in GCTA to estimate the proportion of variance explained by all SNPs [30]. The taxa were divided into two groups based on whether the taxa were present in the ninety percent of participants or not. For alpha diversity and taxa, our model was adjusted for constrain covariates including sex and sequencing batch, as well as quantitative covariates including the first five principal components and age. The model was adjusted for the same covariates except for sequencing batch for analysis of enterotype.

Genetic correlation of gut microbiome and traits

We used GCTA to perform a bivariate GREML analysis to estimate the genetic correlation between gut microbiome and traits in the GNHS[27,37]. The gut microbiome was divided into two groups according to the previous description. We used continuous variables to taxa present at more than ninety percent of participants and traits. For taxa present at fewer than ninety percent of participants, we used binary variables according to the absence/presence of taxa. This analysis included traits such as BMI, FBS, HbA1c, SBP, DBP, HDL-C, LDL-C, TC and TG.

Constructing polygenic scores for taxa and alpha diversity

We selected lead SNPs using PLINK v1.9 with the '—clump' command to clump SNPs that p value $< 5\times10^{-5}$ and $r^2<0.1$ within 0.1 cM. We used beta coefficients as weight to construct polygenic scores for taxa and alpha diversity. For alpha diversity

and taxa present at more than ninety percent participants, we constructed weighted polygenic scores and performed the analysis on a general linear model with a negative binomial distribution to test for association between the polygenic scores and taxa, adjusted for the first five principal components, age, sex and sequencing batch. We used weighted polygenic scores and logistic regression to the absence/presence taxa, adjusted for the same covariates as in the above analysis. Taxa with significance ($p<0.05$) in the replication cohort were included for the further analysis.

The effective number of independent taxa

As some taxa were correlated with each other, we used an eigendecomposition analysis to calculate the effective number of independent taxa on each taxonomy level [35,36]. Matrix M is an m×n matrix, where m is the number of participants and n is the number of taxa on the corresponding taxa level. Matrix A is the variance-covariance matrix of matrix M. P is the matrix of eigenvectors. $\mathrm{diag}\{\lambda_1,\lambda_2,\cdots,\lambda_n\}$ is the diagonal matrix comprised of the ordered eigenvalues, which can be calculated as:

$$\mathrm{diag}\{\lambda_1,\lambda_2,\cdots,\lambda_n\} = P^{-1}AP$$

The effective number of independent taxa can be calculated as:

$$\frac{\left(\sum_{i=1}^{n}\lambda_i\right)^2}{\sum_{i=1}^{n}\lambda_i^2}$$

Bi-directional MR analysis

In the analysis of potential causal effect of gut microbiome features on diseases, we used independent genetic variants (selected as part of the polygenic score analysis) as

the instrument variables. In the analysis of potential causal effect of diseases on gut microbiome features, we selected genetic variants that were replicated in East Asian populations as instrument variables. As all instrument variables were from East Asian populations, we chose independent genetic variants ($r2<0.1$) based on GNHS cohort. We identified the best proxy ($r2 > 0.8$) based on GNHS cohort or discarded the variant if no proxy was available. We used inverse variance weighted (IVW) method to estimate effect size. To confirm the robustness of results, we performed other three MR methods including weighted median, MR-Egger and MR-PRESSO. To assess the presence of horizontal pleiotropy, we performed MR-PRESSO Global test and MR-Egger Regression. Effect sizes of gut microbiome on traits were dependent on units of traits[43] (Supplementary table S1). Results of human complex diseases on the absence/presence gut microbiome were presented as risk of presence (vs absence) of the microbiome per log odds difference of the disease. Results of diseases on other gut microbiome and alpha diversity were presented as changes in abundance of taxa (10-SD of log transformed) per log odds difference of the respective disease.

The statistical significance of gut microbiome on traits and diseases was defined as $p<0.0008$ (0.05/62). In addition, the statistical significance of diseases on gut microbiome features was defined as $p<0.05/n$ (n is the effective number of independent taxa on the corresponding taxonomy level). Results that could not pass Bonferroni adjustment but $p<0.05$ in all four MR methods were considered as potential causal relationships. We performed MR analyses on R v3.5.3.

494

495 **Pathway analysis**

496 We used OTUs by QIIME and annotated the variation of functional genes with
497 Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
498 (PICRUSt) [51]. The pathways and diseases were annotated using KEGG [75-77]. We used
499 Spearman's rank-order correlation to investigate association of predicted pathway or
500 diseases abundance with taxa. In the heatmap, diseases were clustered with 'hcluster'
501 function on R. To test whether non-normalized pathway or disease abundance was
502 associated with each other, we used SPIEC-EASI to test the interaction relationship,
503 and then used Cytoscape v3.7.2 to visualize the interaction network [78,79].

504

505 **Construction of the microbiome risk score**

506 To construct a microbiome risk score for T2DM, we used a Spearman's rank-order
507 correlation to select taxa with the absolute value of correlation coefficient higher than
508 0.2. Score for each taxon abundance <5% quantile in our study was defined as 0. For
509 those above 5%, score for each taxon showing negative association with T2DM was
510 defined as 1; score for each taxon showing positive association with T2DM was
511 defined as -1. We then summed up values from all taxa. We selected logistic
512 regression model to estimate association of the MRS with T2DM risk, and linear
513 model to estimate the association of the MRS with the continuous variables, adjusted
514 for age, sex, dietary energy intake, alcohol intake and BMI at the time of sample
515 collection.

516

517 **Clustering diseases**

518 The clustering analysis was carried out with 'cluster' and 'factoextra' for plot on R.
519 We performed PAM algorithm based on predicted abundance of diseases or average
520 relative abundance after Z-score normalization [80]. PAM algorithm searches k medoids
521 among the observations and then found nearest medoids to minimize the dissimilarity
522 among clusters [18]. Given a set of objects $x = (x_1, x_2,...,x_n)$, the dissimilarity between
523 objects x_i and x_j is denoted by $\mathrm{d}(i,j)$. The assignment of object i to the
524 representative object j is denoted by z_{ij}. z_{ij} is a binary variable and is 1 if object i
525 belongs to the cluster of the representative object j. The function to minimize the
526 model is given by:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d(i,j)z_{ij}$$

527

528 To identify the optimal cluster number, we used 'pamk' function in R to determine the
529 optimum average silhouette width. For each object i, we defined N_i as the average
530 dissimilarity between object i and all other objects within its cluster. For the
531 remaining clusters, $\mathrm{b}(i,w)$ represents the average dissimilarity between i and all
532 objects in cluster w. The minimum dissimilarity M_i can be calculated by:

533 $$M_i = min\forall w(b(i,w)).$$

534 The silhouette width for object i can be calculated by:

$$sw_i = \frac{M_i - N_i}{max(M_i, N_i)}$$

535 Then we calculated the average of silhouette width for each object. The cluster

number is determined by the number of which the average silhouette width is maximum.

Acknowledgments

This study was funded by National Natural Science Foundation of China (81903316, 81773416), Westlake University (101396021801) and the 5010 Program for Clinical Researches (2007032) of the Sun Yat-sen University (Guangzhou, China). The authors declare no conflict of interest. We thank the Westlake University Supercomputer Center for providing computing and data analysis service for the present project.

Data availability

The raw data for 16 S rRNA gene sequences are available in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb at accession number CNP0000829.

References

1. Awany, D. *et al.* Host and Microbiome Genome-Wide Association Studies: Current State and Challenges. *Front Genet* **9**, 637 (2018).
2. Bull, M.J. & Plummer, N.T. Part 1: The Human Gut Microbiome in Health and Disease. *Integr Med (Encinitas)* **13**, 17-22 (2014).
3. Lynch, J.B. & Hsiao, E.Y. Microbiomes as sources of emergent host phenotypes. *Science* **365**, 1405 (2019).
4. Allegretti, J.R., Mullish, B.H., Kelly, C. & Fischer, M. The evolution of the use of faecal microbiota transplantation and emerging therapeutic indications. *The Lancet* **394**, 420-431 (2019).
5. Wong, S.H. & Yu, J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nature Reviews Gastroenterology & Hepatology* (2019).

6. Davies, N.M., Holmes, M.V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).
7. Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet* **48**, 1413-1417 (2016).
8. Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat Genet* **48**, 1396-1406 (2016).
9. Goodrich, J.K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789-99 (2014).
10. Goodrich, J.K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19**, 731-43 (2016).
11. Ganesan, K., Chung, S.K., Vanamala, J. & Xu, B. Causal Relationship between Diet-Induced Gut Microbiota Changes and Diabetes: A Novel Strategy to Transplant Faecalibacterium prausnitzii in Preventing Diabetes. *Int J Mol Sci* **19**(2018).
12. He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med* **24**, 1532-1535 (2018).
13. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210-215 (2018).
14. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A. & Alm, E.J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications* **8**, 1784-1784 (2017).
15. Cheng, S. *et al.* Identifying psychiatric disorder-associated gut microbiota using microbiota-related gene set enrichment analysis. *Briefings in Bioinformatics* (2019).
16. Jackson, M.A. *et al.* Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nature Communications* **9**, 2655 (2018).
17. Cao, Y. *et al.* Association of magnesium in serum and urine with carotid intima-media thickness and serum lipids in middle-aged and elderly Chinese: a community-based cross-sectional study. *European journal of nutrition* **55**(2015).
18. Kaufman, L. & Rousseeuw, P. Partitioning Around Medoids (Program PAM). 68-125 (1990).
19. Sun, L.-L. *et al.* Associations between the dietary intake of antioxidant nutrients and the risk of hip fracture in elderly Chinese: A case-control study. *The British journal of nutrition* **112**, 1-9 (2014).
20. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575 (2007).
21. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* **48**, 1284 (2016).
22. Clarke, L. *et al.* The international Genome sample resource (IGSR): A

worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Research* **45**, D854-D859 (2016).

23. Delaneau, O. *et al.* Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications* **5**, 3934 (2014).
24. Okada, Y. *et al.* Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. *Hum Mol Genet* **23**, 6916-26 (2014).
25. Pillai, N.E. *et al.* Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. *Hum Mol Genet* **23**, 4443-51 (2014).
26. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLOS ONE* **8**, e64683 (2013).
27. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
28. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. & Price, A.L. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**, 100-6 (2014).
29. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174-80 (2011).
30. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).
31. Hua, X. *et al.* MicrobiomeGWAS: a tool for identifying host genetic variants associated with microbiome composition. *bioRxiv*, 031187 (2015).
32. Ruhlemann, M.C. *et al.* Application of the distance-based F test in an mGWAS investigating beta diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci. *Gut Microbes* **9**, 68-75 (2018).
33. Bonder, M.J. *et al.* The effect of host genetics on the gut microbiome. *Nat Genet* **48**, 1407-1412 (2016).
34. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565 (2010).
35. Bretherton, C.S., Widmann, M., Dymnikov, V.P., Wallace, J.M. & Bladé, I. The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field. *Journal of Climate* **12**, 1990-2009 (1999).
36. Wang, H. *et al.* Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Science Advances* **5**, eaaw3538 (2019).
37. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2 (2012).
38. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics* **51**, 600-605 (2019).
39. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread

horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* **50**, 693-698 (2018).

40. Low, S.K. *et al.* Identification of six new genetic loci associated with atrial fibrillation in the Japanese population. *Nat Genet* **49**, 953-958 (2017).
41. Suzuki, K. *et al.* Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat Genet* **51**, 379-386 (2019).
42. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat Genet* **49**, 1458-1467 (2017).
43. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat Genet* **50**, 390-400 (2018).
44. Matoba, N. *et al.* GWAS of smoking behaviour in 165,436 Japanese people reveals seven new loci and shared genetic architecture. *Nat Hum Behav* **3**, 471-477 (2019).
45. Lu, X.F. *et al.* Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. *Nature Genetics* **44**, 890-+ (2012).
46. Marzec, J. *et al.* A genetic study and meta-analysis of the genetic predisposition of prostate cancer in a Chinese population. *Oncotarget* **7**, 21393-403 (2016).
47. Okada, Y. *et al.* Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat Genet* **44**, 904-9 (2012).
48. Zeng, C. *et al.* Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology* **150**, 1633-1645 (2016).
49. Zhou, X. *et al.* Identification of genetic risk factors in the Chinese population implicates a role of immune system in Alzheimer's disease pathogenesis. *Proc Natl Acad Sci U S A* **115**, 1697-1706 (2018).
50. Foo, J.N. *et al.* Genome-wide association study of Parkinson's disease in East Asians. *Hum Mol Genet* **26**, 226-232 (2017).
51. Langille, M.G.I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**, 814 (2013).
52. Canga, Y. *et al.* Assessment of Atrial Conduction Times in Patients with Newly Diagnosed Parkinson's Disease. *Parkinsons Dis* **2018**, 2916905 (2018).
53. Ihara, M. & Washida, K. Linking Atrial Fibrillation with Alzheimer's Disease: Epidemiological, Pathological, and Mechanistic Evidence. *J Alzheimers Dis* **62**, 61-72 (2018).
54. Conen, D. *et al.* Risk of Malignant Cancer Among Women With New-Onset Atrial FibrillationAtrial Fibrillation and Risk of CancerAtrial Fibrillation and Risk of Cancer. *JAMA Cardiology* **1**, 389-396 (2016).
55. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53-65 (1987).

56. Hu, F.B. *et al.* Prospective Study of Cataract Extraction and Risk of Coronary Heart Disease in Women. *American Journal of Epidemiology* **153**, 875-881 (2001).
57. Javaheri, S. & Redline, S. Insomnia and Risk of Cardiovascular Disease. *Chest* **152**, 435-444 (2017).
58. Strand, L.B. *et al.* Self-reported sleep duration and coronary heart disease mortality: A large cohort study of 400,000 Taiwanese adults. *International Journal of Cardiology* **207**, 246-251 (2016).
59. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* **16**, 191 (2015).
60. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415-25 (2010).
61. Vinter, N., Christesen Amanda, M.S., Fenger-Grøn, M., Tjønneland, A. & Frost, L. Atrial Fibrillation and Risk of Cancer: A Danish Population-Based Cohort Study. *Journal of the American Heart Association* **7**, e009543 (2018).
62. Zoja, C. *et al.* Imatinib ameliorates renal disease and survival in murine lupus autoimmune disease. *Kidney International* **70**, 97-103 (2006).
63. Boursi, B., Mamtani, R., Haynes, K. & Yang, Y.-X. Parkinson's disease and colorectal cancer risk-A nested case control study. *Cancer epidemiology* **43**, 9-14 (2016).
64. Xie, X., Luo, X. & Xie, M. Association between Parkinson's disease and risk of colorectal cancer. *Parkinsonism & Related Disorders* **35**, 42-47 (2017).
65. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet* **48**, 1043-8 (2016).
66. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-8 (2013).
67. Pankratz, N. *et al.* Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2. *Annals of neurology* **71**, 370-384 (2012).
68. Anderson, C.A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-73 (2010).
69. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179 (2011).
70. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).
71. Edgar, R.C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**, 996-8 (2013).
72. Caporaso, J.G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-6 (2010).
73. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1-27 (1974).
74. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**,

291-295 (2015).

75. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

76. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res* **47**, D590-D595 (2019).

77. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* (2019).

78. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).

79. Kurtz, Z.D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology* **11**, e1004226 (2015).

80. Reynolds, A.P., Richards, G., de la Iglesia, B. & Rayward-Smith, V.J. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms* **5**, 475-504 (2006).

Figure legends

Figure 1 Association of host polygenic score with gut microbiome. The participants were divided into high and low polygenic score group according to median levels of the polygenic score. The dots on the right of the box represent the distribution of polygenic score. The dash line in the box is the position of median line and the solid line is the position of mean line. The length of box depends on upper quartile and lower quartile of datum. Sample size at the discovery stage is 1475, and that at replication stage is 199. **(A).** Correlation of *Coriobacteriaceae* abundance with the polygenic score (including 45 lead SNPs, Supplementary Table S8). **(B).** Correlation of undefined species belonging to *Parabacteroides* genus (*Parabacteroides_undefined*) with the polygenic score (including 32 lead SNPs, Supplementary Table S8). **(C).** Correlation of *Odoribacter* presence with the polygenic score (including 43 lead SNPs, Supplementary Table S8).

Figure 2 Effect of host genetically predicted higher atrial fibrillation risk on gut microbiome. (A). Causal association of atrial fibrillation with abundance of *Burkholderiales*, *Alcaligenaceae*, *Lachnobacterium* and *Coprophilus*. The effect sizes of atrial fibrillation on taxa are changes in abundance of bacteria (10-SD of log-transformed) per genetically determined higher log odds of atrial fibrillation. **(B)**. Causal association of atrial fibrillation with presence of *Barnesiellaceae*, *Veillonellaceae_undefined* and *Mitsuokella*. The effect size of atrial fibrillation on taxa are present as odds ratio increase in log odds of atrial fibrillation. **(C)**. The heat map shows correlation of AF-associated taxa with predicted diseases. The grey components show no significance of correlation with Bonferroni correction ($p>0.05/(5.6*22)$, $p>0.0004$).

Figure 3 Association and cluster of diseases predicted by the gut microbiome. (A). Plot of clusters in Guangzhou Nutrition and Health Study (GNHS) cohort (n=1919). **(B)**. Plot of cluster results in the replication cohort (n=217). **(C)**. Plot of 5 clusters in antibiotic-taking participants (n=18). The optimal cluster is 5 in GNHS cohort and 6 in the replication. The clusters share consistent components between two studies. In contrast, components are different between antibiotic-taking participants and control groups. Dimension1 (Dim1) and dimension2 (Dim2) can explain 40.1% and 13.1% variance, respectively in GNHS cohort. The annotation for variables is as following. AT: African trypanosomiasis, AD: Alzheimer's disease, V1: Amoebiasis, ALS: Amyotrophic lateral sclerosis, BC: Bladder cancer, CD: Chagas disease, CML: Chronic myeloid leukemia, CRC: Colorectal cancer, V2: Hepatitis C, HD: Huntington's disease, HCM: Hypertrophic cardiomyopathy, V3: Influenza A, PD: Parkinson's disease, V4: Pathways in cancer, V5: Prion disease, PCa: Prostate cancer, RCC: Renal cell carcinoma, SLE: Systemic lupus erythematosus, V6: Tuberculosis, T1DM: Type I diabetes mellitus, T2DM: Type II diabetes mellitus, V7: Vibrio cholerae infection. **(D)**. Plot of clusters in GNHS patients. Patients get only one of the follow diseases: stroke (n=8), chronic hepatitis (n=19), coronary heart diseases (n=40),

cataract (n=124) and insomnia (n=68). **(E)**. Gut microbiome-predicted network of relationship among different human complex diseases. The interaction is determined by SPIEC-EASI with non-normalized predicted abundance data.

Figure 4 Correlation of the human complex diseases with gut microbiome on genus level. The heat map shows correlation of predicted diseases and gut microbiome on genus level. The grey components show no significance of correlation with Bonferroni correction ($p>0.05/ (5.6*22)$, $p>0.0004$).

Figure 1 Association of host polygenic score with gut microbiome. The participants were divided into high and low polygenic score group according to median levels of the polygenic score. The dots on the right of the box represent the distribution of polygenic score. The dash line in the box is the position of median line and the solid line is the position of mean line. The length of box depends on upper quartile and lower quartile of datum. Sample size at the discovery stage is 1475, and that at replication stage is 199. **(A).** Correlation of *Coriobacteriaceae* abundance with the polygenic score (including 45 lead SNPs, Supplementary Table S8). **(B).** Correlation of undefined species belonging to *Parabacteroides* genus (*Parabacteroides_undefined*) with the polygenic score (including 32 lead SNPs, Supplementary Table S8). **(C).** Correlation of *Odoribacter* presence with the polygenic score (including 43 lead SNPs, Supplementary Table S8).

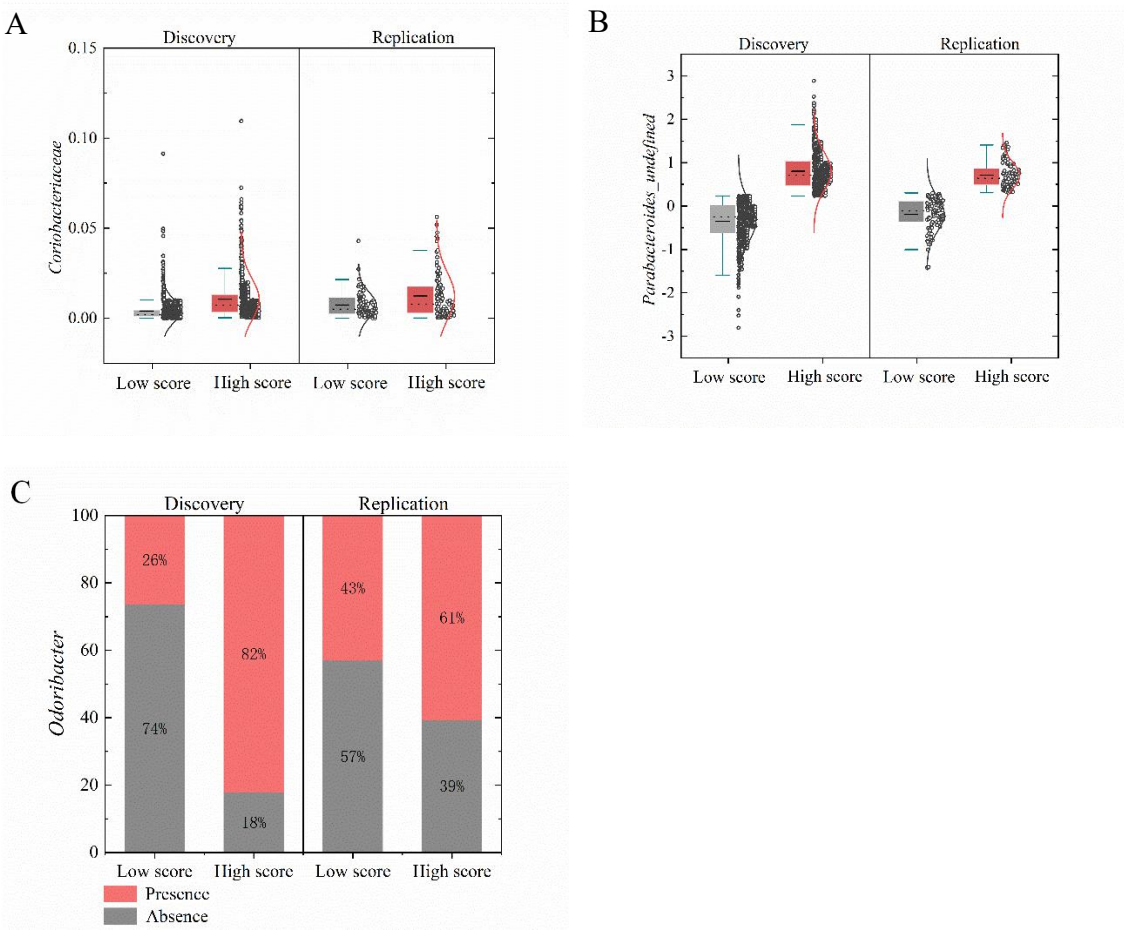
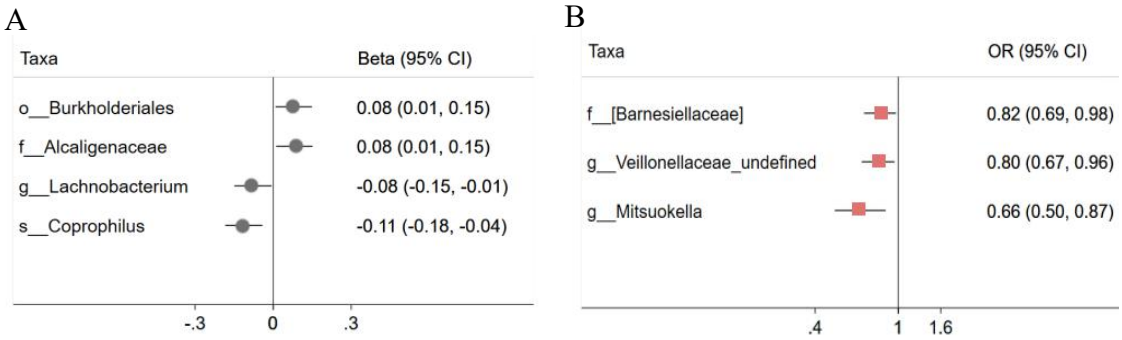


Figure 2 Effect of host genetically predicted higher atrial fibrillation risk on gut microbiome. (A). Causal association of atrial fibrillation with abundance of *Burkholderiales*, *Alcaligenaceae*, *Lachnobacterium* and *Coprophilus*. The effect sizes of atrial fibrillation on taxa are changes in abundance of bacteria (10-SD of log-transformed) per genetically determined higher log odds of atrial fibrillation. **(B)**. Causal association of atrial fibrillation with presence of *Barnesiellaceae*, *Veillonellaceae_undefined* and *Mitsuokella*. The effect size of atrial fibrillation on taxa are present as odds ratio increase in log odds of atrial fibrillation. **(C)**. The heat map shows correlation of AF-associated taxa with predicted diseases. The grey components show no significance of correlation with Bonferroni correction (p>0.05/(5.6*22), p>0.0004).



C

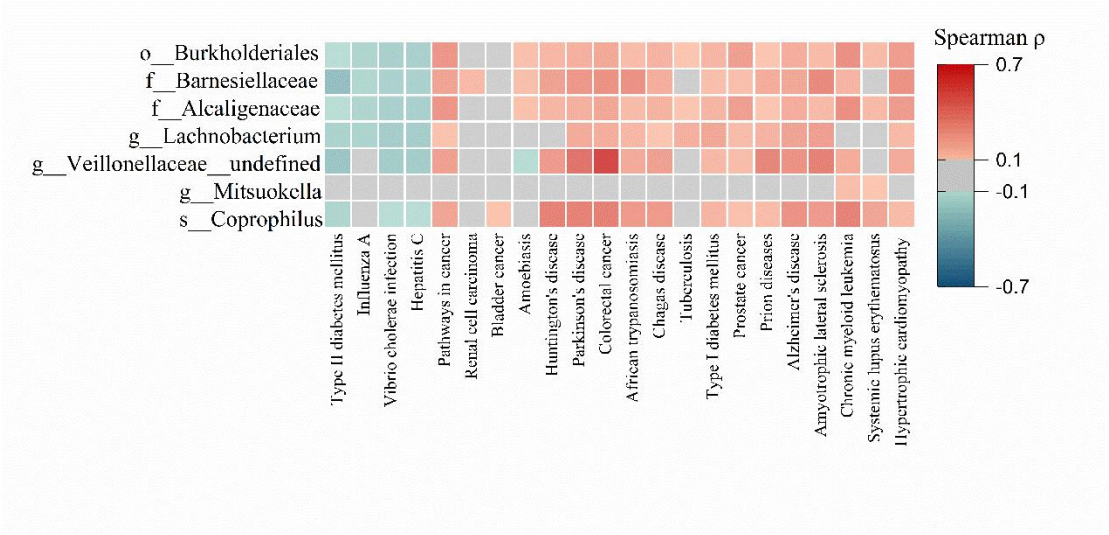
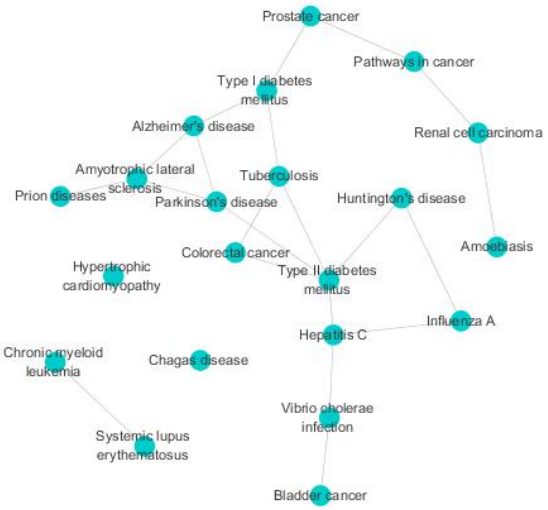


Figure 3 Association and cluster of diseases predicted by the gut microbiome. (A). Plot of clusters in Guangzhou Nutrition and Health Study (GNHS) cohort (n=1919). **(B)**. Plot of cluster results in the replication cohort (n=217). **(C)**. Plot of 5 clusters in antibiotic-taking participants (n=18). The optimal cluster is 5 in GNHS cohort and 6 in the replication. The clusters share consistent components between two studies. In contrast, components are different between antibiotic-taking participants and control groups. Dimension1 (Dim1) and dimension2 (Dim2) can explain 40.1% and 13.1% variance, respectively in GNHS cohort. The annotation for variables is as following. AT: African trypanosomiasis, AD: Alzheimer's disease, V1: Amoebiasis, ALS: Amyotrophic lateral sclerosis, BC: Bladder cancer, CD: Chagas disease, CML: Chronic myeloid leukemia, CRC: Colorectal cancer, V2: Hepatitis C, HD: Huntington's disease, HCM: Hypertrophic cardiomyopathy, V3: Influenza A, PD: Parkinson's disease, V4: Pathways in cancer, V5: Prion disease, PCa: Prostate cancer, RCC: Renal cell carcinoma, SLE: Systemic lupus erythematosus, V6: Tuberculosis, T1DM: Type I diabetes mellitus, T2DM: Type II diabetes mellitus, V7: Vibrio cholerae infection. **(D)**. Plot of clusters in GNHS patients. Patients get only one of the follow diseases: stroke (n=8), chronic hepatitis (n=19), coronary heart diseases (n=40), cataract (n=124) and insomnia (n=68). **(E)**. Gut microbiome-predicted network of relationship among different human complex diseases. The interaction is determined by SPIEC-EASI with non-normalized predicted abundance data.

E



862

863

Figure 4 Correlation of the human complex diseases with gut microbiome. The heat map shows correlation of predicted diseases and gut microbiome on genus level. The grey components show no significance of correlation with Bonferroni correction (p>0.05/ (5.6*22), p>0.0004).