

ARTICLE OPEN



Chinese gut microbiota and its associations with staple food type, ethnicity, and urbanization

Jing Lu^{1,2,9}, Li Zhang^{3,9}, Qixiao Zhai^{1,2,4}, Jianxin Zhao^{1,2,4}, Hao Zhang^{1,2,4}, Yuan-Kun Lee^{5,6}, Wenwei Lu^{1,2,4,5}✉, Mingkun Li^{3,7,8}✉ and Wei Chen^{1,2,4,5}✉

The gut microbiota could affect human health and disease. Although disease-associated microbiota alteration has been extensively investigated in the Chinese population, a nationwide Chinese gut microbiota baseline is still lacking. Here we performed 16 S rRNA gene sequencing on fecal samples from 2678 healthy Chinese individuals, who belonged to eight ethnic groups and resided in 63 counties/cities of 28 provinces. We identified four enterotypes, three of which were enriched for *Prevotella*, *Bacteroides*, and *Escherichia*, respectively, whereas the fourth one had no dominant genus. By assessing the association between the gut microbiota and 20 variables belonging to six categories, geography, demography, diet, urbanization, lifestyle, and sampling month, we revealed that geography explained the largest microbiota variation, and clarified the distinct patterns in the associations with staple food type, ethnicity, and urban/rural residence. Specifically, the gut microbiota of Han Chinese and ethnic minority groups from the same sites was more alike than that of the same ethnic minority groups from different sites. Individuals consuming wheat as staple food were predicted to have more microbial genes involving in glucan 1,3-beta-glucosidase and S-adenosyl-L-methionine biosynthesis than those who consumed rice, based on functional prediction. Besides, an appreciable effect of urbanization on decreased intra-individual diversity, increased inter-individual diversity, and increased proportion of the *Bacteroides* enterotype was observed. Collectively, our study provided a nationwide gut microbiota baseline of the Chinese population and knowledge on important covariates, which are fundamental to translational microbiota research.

npj Biofilms and Microbiomes (2021)7:71; <https://doi.org/10.1038/s41522-021-00245-0>

INTRODUCTION

The human intestine harbors a special diverse microbial ecosystem, with an estimated 150–400 bacterial species reside in our gut¹. The gut microbiota provides substantial benefits to our health by forming a barrier against pathogens, producing bioactive metabolites, and regulating immunological functions. The homeostasis of the gut ecosystem is maintained by some core species that are generally shared among different individuals, and the gut microbiota in healthy adults is relatively stable in the absence of strong influencing factors (e.g., dietary changes or antibiotic treatment)^{2,3}.

The imbalance of the gut microbiota (i.e., dysbiosis) is associated with many diseases, e.g., inflammatory bowel disease, obesity, allergies, and autoimmune diseases⁴. Plenty of microbial components have been revealed to involve in a series of pathologies by extensive disease-targeted microbiota researches, and can thus, in theory, serve as biomarkers. For example, fecal microbial markers for screening colorectal cancer have been widely studied⁵. However, the translation of microbiota research into clinical practice is still limited by multiple challenges, especially the difficulty in the precise classification of “healthy” microbiota, which requires comprehensive knowledge of the microbiota variation and covariates of an average, healthy population. A study based on 7009 individuals from 14 geographic districts in one province of China demonstrated that microbiota-based metabolic disease models developed in one location could

not be extrapolated to other locations, and the efficiency of interpolated models decreased as geographic scale increased⁶. This emphasized the influence of geography on gut microbiota composition and disease model application, whereas the microbiota variation on a larger geographic scale across China is yet to be explored.

Besides geography, many gut microbiota covariates have been uncovered, including diet, lifestyle, ethnicity, socioeconomic status, medication, and genetics^{7,8}. For example, stratification of the gut microbiota (termed enterotypes) has been associated with diet, especially the intake of fibers and carbohydrates⁹; urbanization has been associated with increased inter-individual variation and loss of species with high potential for fiber degradation¹⁰; taxa that were differentially abundant across ethnicities have been proposed to be associated with chronic diseases^{11,12}. To date, most population-level studies investigating gut microbiota covariates have focused on the western population (Europe and the USA) and a few in Israel, Japan, and China¹³, but rarely in Africa, South America, and other regions of Asia.

There are various food styles in China, and people living in different regions show great varieties in their diet. Moreover, there are 56 ethnic groups in China, which have distinct characteristics in diet, lifestyle, custom, and culture. Therefore, the diversified gut microbiota is expected in the Chinese population. Meanwhile, as a result of an unprecedented speed and scale of urbanization, Chinese is undergoing rapid change in lifestyle, and the dietary

¹State Key Laboratory of Food Science and Technology, Jiangnan University, Wuxi, China. ²School of Food Science and Technology, Jiangnan University, Wuxi, China. ³Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation, Beijing, China. ⁴National Engineering Research Center for Functional Food, Jiangnan University, Wuxi, China. ⁵International Joint Research Laboratory for Pharmabiotics & Antibiotic Resistance, Jiangnan University, Wuxi, China. ⁶Department of Microbiology & Immunology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. ⁷University of Chinese Academy of Sciences, Beijing, China. ⁸Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. ⁹These authors contributed equally: Jing Lu, Li Zhang. ✉email: luwenwei@jiangnan.edu.cn; limk@big.ac.cn; chenwei66@jiangnan.edu.cn

habit is shifting towards a western-style diet. Specifically, more high-fat and high-protein foods are consumed while fewer grains are taken, which might significantly change the gut microbiota as observed in other developing countries. Although a couple of studies have been conducted to investigate the gut microbiota characteristics in China, these studies either focused on limited regions or recruited a small number of participants, and a nationwide gut microbiota survey is still missing^{6,14,15}.

To characterize the gut microbiota diversity in the Chinese population and investigate microbiota-associated variables, we collected feces of 2678 individuals without apparent diseases (referred to as “healthy”), which underwent 16 S ribosomal RNA (rRNA) gene sequencing (V3–4 region). A questionnaire including information on demography, diet, and lifestyle was carried out, which enables an in-depth analysis of factors associated with the Chinese gut microbiota.

RESULTS

Overview of cohort and data

We recruited 2678 healthy volunteers (male 1144, female 973) from 63 counties/cities of 28 provinces, including 2167 Han Chinese (1755 with age over 3 and 412 with age under 3), 487 individuals from seven ethnic minority groups (Tibetan, 156; Hui, 107; Miao, 73; Uyгур, 70; Naxi, 46; Mongolian, 41; Bai, 18), and 24 individuals without ethnicity information (Fig. 1a, b). Fecal samples were collected following a standardized procedure (see Methods for details). Meanwhile, 20 phenotypical and environmental variables were collected via questionnaires or national annals, and classified into six categories: geography, demography, diet, urbanization, lifestyle, and sampling month (Supplementary Data 1). The gut microbiota was profiled by sequencing the variable region 3–4 (V3–4) of 16 S rRNA gene, with a median read number of 27,638 per sample (range 10,000–236,350). The reads were clustered into 14,364 zero-radius operational taxonomic units (ZOTUs), and 56.64% of these ZOTUs (accounting for 86.48% of the total reads) were assigned to 444 genera belonging to 24 phyla.

It is known that a stable gut microbiota resembling that of adults is established at age 3¹⁶. In our data, we observed a strong positive correlation between the alpha diversity and age in children under age 3 (Shannon index, $R^2 = 0.37$, observed ZOTUs, $R^2 = 0.31$, Faith’s phylogenetic diversity (Faith’s PD), $R^2 = 0.35$, $p < 2.2 \times 10^{-16}$), but not in other age groups (age 3–17, 18–65 or 66–112, $p > 0.05$, Supplementary Fig. 1). Thus, only 2266 individuals with ages 3–112 (median 46) were included in the following analyses.

The gut microbiota composition of the Chinese population and associated covariates

Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria were the four most abundant bacterial phyla in all samples (Fig. 2a, b). A total of 24 genera were observed in >90% of samples with average relative abundances >0.1% (the core microbiota, Fig. 2b). Eighteen of these genera overlapped with the core gut microbiota of 2008 healthy Chinese individuals who resided in Guangdong province⁶ (Supplementary Data 2); seven of them overlapped with the top nine most abundant fecal genera in another Chinese cohort, which included 314 healthy individuals from nine provinces¹⁴; ten of them overlapped with the top 20 fecal genera discovered by the Human Microbiome Project¹⁷. We further stratified the microbiota into four enterotypes using the clustering method described by Arumugam et al.¹⁸. We identified driving genera by random forest algorithm (area under the curve (AUC) for receiver operating characteristic (ROC) curve: 0.99, Supplementary Fig. 2a), obtaining *Prevotella* enterotype (E1, $n = 443$), *Bacteroides* enterotype (E2, $n = 732$), *Escherichia* enterotype (E3, $n = 251$), and mixture enterotype (E4, $n = 840$) (Fig. 2c, d). E1

and E2 are two well-recognized enterotypes, whereas unlike Firmicutes (most prominently *Ruminococcus*) being the third enterotype in most previous studies¹⁹, E3 was distinguished by an overrepresentation of *Escherichia* (a genus belonging to family Enterobacteriaceae, phylum Proteobacteria), which has rarely been reported²⁰. Besides, E4 showed no dominant genus but a mixture of a few relatively abundant genera, including *Bifidobacterium* and *Blautia* etc.

Covariates associated with the alpha diversity of gut microbiota were first investigated with simple linear regressions. Faith’s PD and Observed ZOTUs were significantly correlated with geographical zone (including ten zones differing in climate, topography, etc., Supplementary Fig. 2b), altitude, staple food type, urban/rural residence, ethnicity, and sampling month (adjusted $R^2 > 0.01$, $p < 0.0001$, Fig. 2e). The correlations were validated in multiple linear models incorporating all six covariates, except that with altitude (Supplementary Data 3). This reflected the dependence of altitude, which was different among ethnic groups (Supplementary Data 1). In addition, a simple linear model using only Han individuals did not support the correlation between alpha diversity and altitude either.

Meanwhile, the gut microbial community structure (beta diversity, estimated by Jensen-Shannon divergence (JSD)) was significantly correlated with 17 covariates as evaluated by envfit²¹ ($p_{\text{adj}} < 0.05$, Fig. 2f). Geographic factors (sampling site, province, latitude, and geographical zone) explained the largest variance, followed by age, sampling month, ethnicity, staple food type, urbanization, and other geographic factors (Supplementary Fig. 2c, d for Bray–Curtis and unweighted UniFrac distances). To further explore the correlation between geographic location and the gut microbiota, we applied the Mantel test on the microbial JSD matrix and geographic distance matrix, and found a significant correlation between them ($p = 0.03$, Supplementary Fig. 2e), suggesting that the gut microbiota change gradually in proximal locations. Detailed analysis of these associations was conducted in the following sections.

Association between staple food type and the gut microbiota

Samples were assigned into three groups according to the dominant staple food regularly taken, i.e., rice (white rice), wheat (white flour of common wheat), and rice & wheat. Owing to the requirements in temperature, precipitation, and sunshine duration for different grain crops, wheat was mainly grown in northern China, whereas rice was cultivated more widely²². Consumption of the two grains also showed similar geographic distribution (Fig. 3a). Intriguingly, the alpha diversity indices including Faith’s PD, Shannon index and Observed ZOTUs were significantly higher in individuals/regions consuming more rice ($p < 0.01$, Fig. 3a, Supplementary Fig. 3a–c).

Bacterial genera differing between each two groups were identified using DESeq2 analysis with adjustment for age and gender ($p_{\text{adj}} < 0.05$, Fig. 3b). *Bifidobacterium* and *Catenibacterium* were enriched in individuals consuming wheat and wheat & rice when comparing to individuals consuming rice. The abundances of these two genera decreased gradiently in the three groups, indicating that there might be a dose-effect associated with wheat intake. The association between *Bifidobacterium* and wheat intake is consistent with previous observations that the abundance of *Bifidobacterium* was decreased when taking diets with low wheat content, e.g., gluten-free diet, low-gluten diet, and low FODMAP (fermentable oligosaccharides, disaccharides, monosaccharides, and polyols) diet^{23–28}. Twelve genera were enriched in individuals consuming rice and rice & wheat when comparing to individuals consuming wheat, of which *Bacteroides*, *Parabacteroides*, a butyrate producer *Clostridium XIVA* and an opportunistic pathogen *Fusobacterium* were the most significant ones ($\log_2\text{FC} > 1$, $p_{\text{adj}} <$

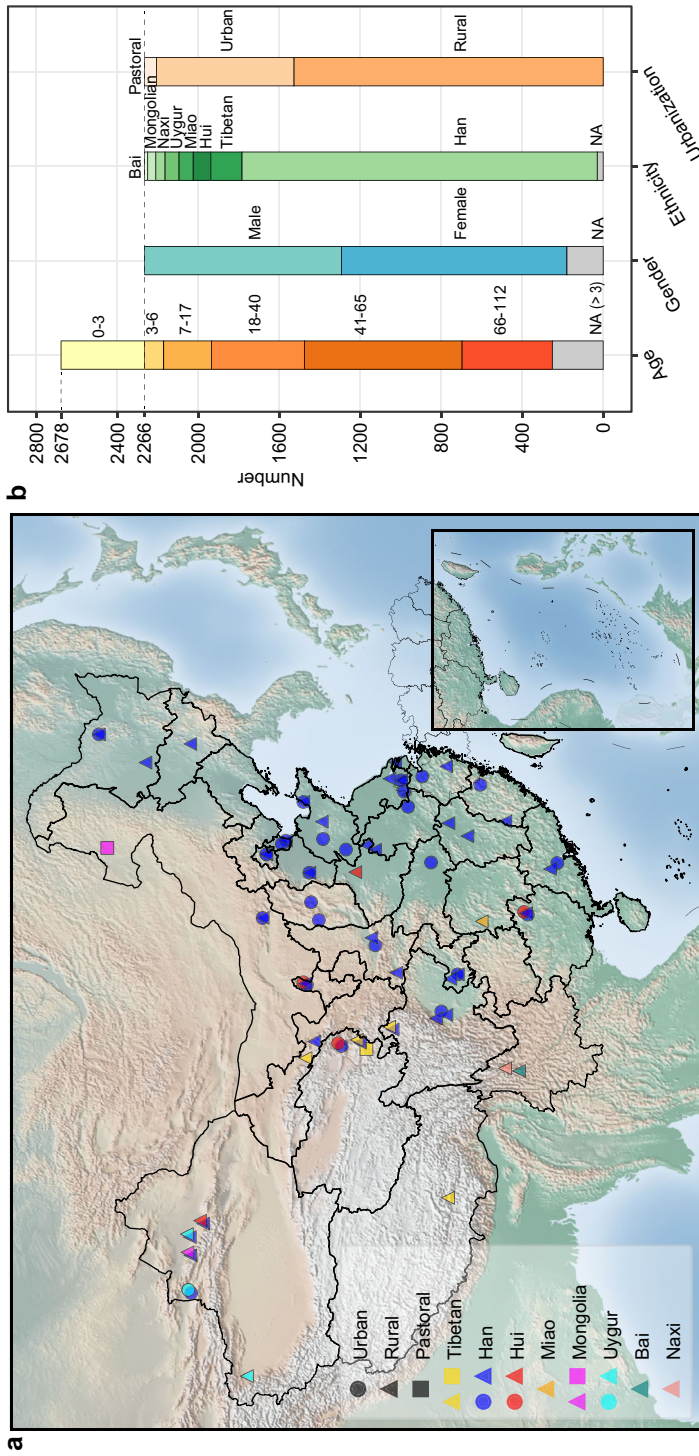


Fig. 1 Overview of cohort. **a** Geographic distribution of sampling sites, ethnic groups, and urbanization status. Sampling sites were mapped using the Matplotlib Basemap Toolkit. **b** Cohort composition with respect to age, gender, ethnicity, and urbanization status. See also Figure S1.

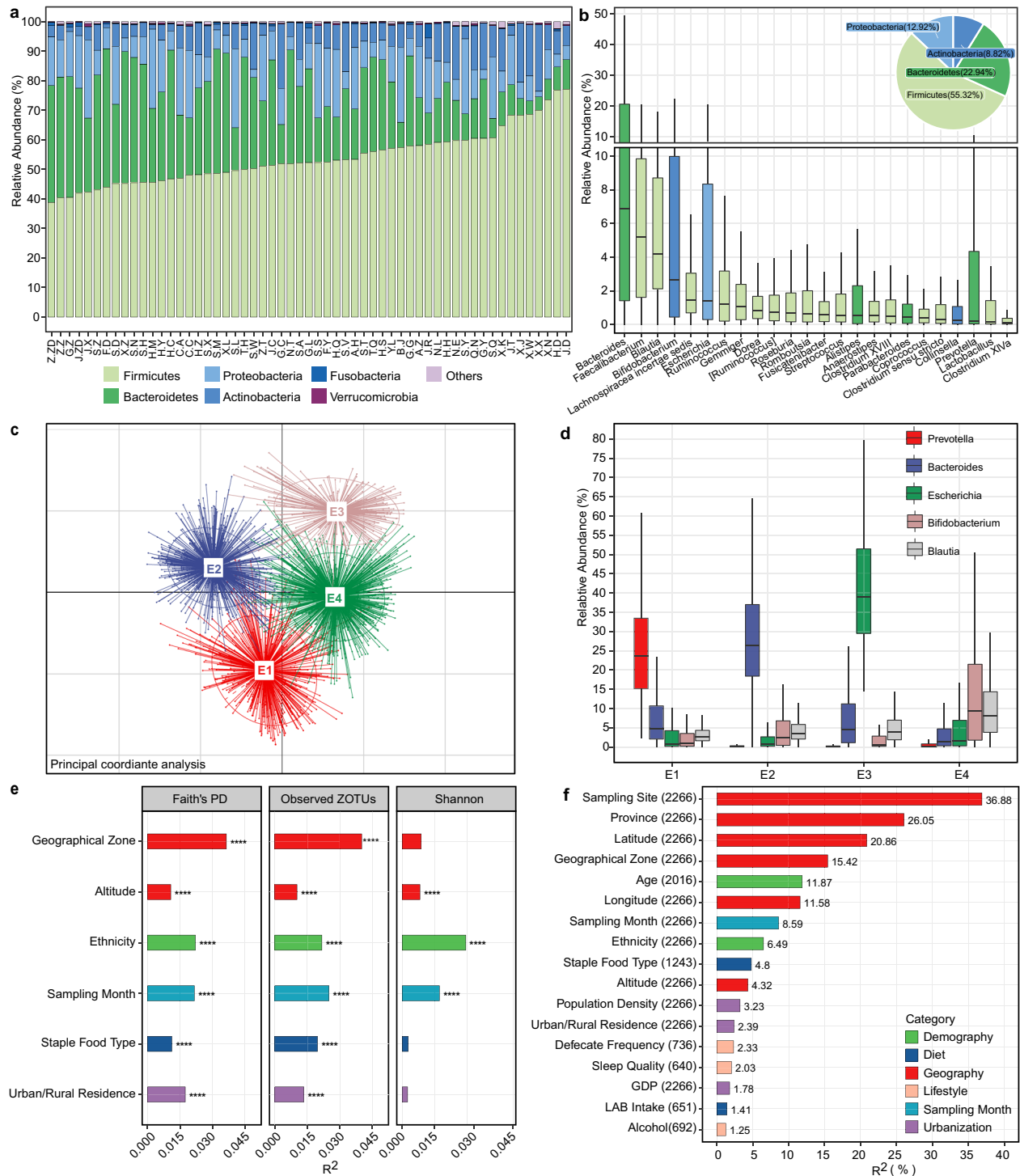


Fig. 2 The gut microbiota composition of the Chinese population and associated covariates. **a** Relative abundances of the top six phyla at each sampling site. **b** Relative abundances of the 24 core genera, which presented in >90% of individuals with an average relative abundance >0.1%. The pie chart shows the phylum-level microbial composition of the cohort. LAB: lactic acid bacteria. **c** PCoA plot showing four enterotypes. **d** Relative abundances of representative genera of enterotypes. **e** Covariates associated with microbiota alpha diversity. Covariates having adjusted $R^2 > 0.01$ and **** $p < 0.0001$ (simple linear regression) with at least one alpha index were shown. **f** Covariates associated with microbiota beta diversity estimated by JSD distance. The effect size was calculated with envfit (vegan), and covariates with $p_{adj} < 0.05$ were shown. The number of samples was indicated in brackets following each covariate. In boxplots, the center line represents the median, box limits represent upper and lower quartiles and whiskers represent 1.5 \times interquartile range. See also Figure S2.

1e-10). However, no difference was observed in the enterotype composition among the three groups.

To explore the metabolism capacity of gut microbiota affected by distinct staple foods, enzyme commission (EC) numbers and

MetaCyc pathways were inferred with PICRUSt2²⁹. Considering that common wheat flour contains more dietary fiber than white rice (2–3% vs. 0.7–2% of dry matter) and the fiber components differ significantly between the two grains^{30,31}, we specifically

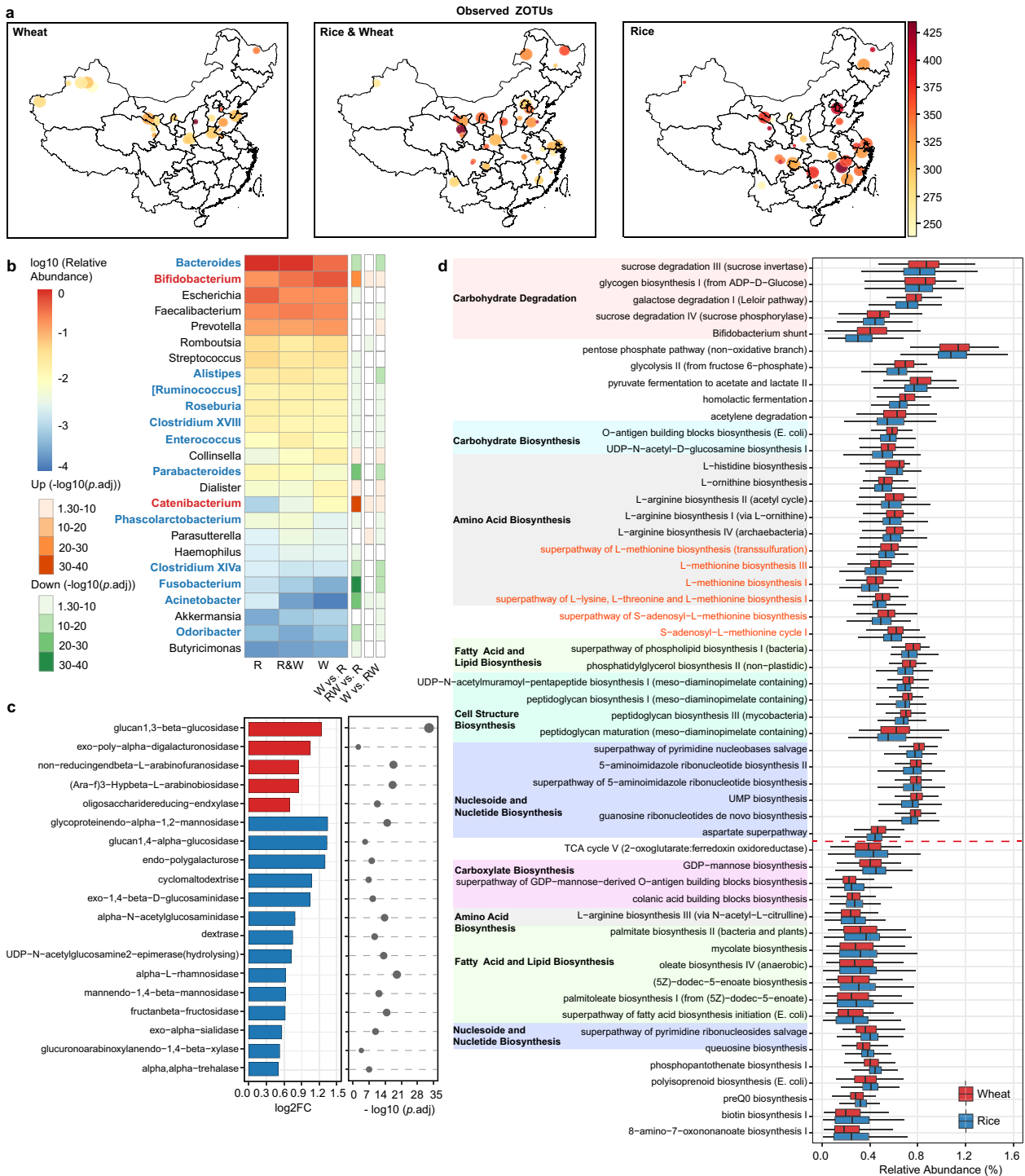


Fig. 3 Differences in the microbiota alpha diversity, composition, and functional potential between populations consuming different staple food. **a** Observed ZOTUs. Each dot represents a sampling site; the color of the dot indicates the median value in each site; the diameter of the dot is proportional to the number of samples in each site and is fixed for sites with >15 samples. **b** Differential genera between the three groups. Genera with $p.adj < 0.05$ from DESeq2 analyses were shown. Genera highlighted in red or blue indicate higher abundances associated with wheat or rice intake, which were detected with two DESeq2 models. *R* rice, *W* wheat. **c** Differential glycosidase between the Rice and Wheat group. Glycosidase with $\log_2FC > 0.5$ and $p.adj < 0.05$ from Mann–Whitney tests were shown. **d** Differential pathways between the Rice and Wheat group. Pathways were grouped by the functional category, and those with higher abundance in the Wheat group were placed on the upper part of the plot. Pathways related to L-methionine and S-adenosyl-L-methionine biosynthesis were highlighted in red. Pathways with $\log_2FC > 0.1$ and $p.adj < 1e-10$ from Mann–Whitney tests were shown. In boxplots, the center line represents the median, box limits represent upper and lower quartiles and whiskers represent 1.5× interquartile range. Rice: $n = 417$, Wheat: $n = 549$, Rice & Wheat: $n = 277$. See also Figure S3.

focused on 69 ECs belonging to glycosidase (EC 3.2.1). The abundances of 19 glycosidases differed between the Wheat and Rice group ($\log_2FC > 0.5$, $p_{\text{adj}} < 0.05$, Fig. 3c). The Wheat group showed dramatically higher glucan 1,3-beta-glucosidase (EC 3.2.1.58), in line with the fact that beta-glucan containing beta-(1-3)-linkages exists in wheat but not rice. A total of 53 pathways differed moderately between the Wheat and Rice group ($\log_2FC > 0.1$, $p_{\text{adj}} < 1e-10$, Fig. 3d). First, the Wheat group was distinguished by the higher potential of a few carbohydrate degradation pathways, as well as glycolysis, pentose phosphate pathway, and lactate/acetate fermentation. Second, the Wheat group showed increased capacity for biosynthesis of amino acids including L-methionine, S-adenosyl-L-methionine (SAM) and L-arginine etc. Of note, SAM is widely adopted as a therapy for liver disease, depression, and osteoarthritis^{32,33}. Besides, the Wheat group was associated with a higher potential of housekeeping functions including cell structure biosynthesis and nucleic acid processing. These findings indicate that the staple food type and possibly related dietary habits may alter the metabolism capacity of gut microbiota.

Association between ethnicity and the gut microbiota

Among eight ethnic groups included in this study, Tibetan had the highest alpha diversity, whereas Bai had the lowest alpha diversity (Supplementary Fig. 4a). In addition, the gut microbial community structure differed between ethnic groups ($R^2 = 4.00\%$, $p < 0.001$, permutational multivariate analysis of variance (PERMANOVA) based on JSD, supplementary Fig. 4b). However, as some ethnic groups reside in specific geographic locations, it was hard to partial out the geographic effect. In our study, four ethnic minority groups, namely, Uygur, Hui, Mongolian, and Tibetan, each had not only samples collected from different sites, but also accompanying Han Chinese samples collected from the same sites, which enabled us to distinguish the effect of ethnicity on microbiota from that of geography. The gut microbiota richness (Observed ZOTUs) differed between samples belonging to the same ethnic groups but from different sites (at least 200 km apart) for all four ethnic minority groups. In contrast, the microbiota richness of different ethnic groups from the same sites did not show significant differences except that between Uygur and Han ($p_{\text{adj}} < 0.05$, Fig. 4a, supplementary Fig. 4c, d for Shannon index and Faith's PD). As to microbiota beta diversity, clustering by sampling site and ethnic group were both distinguishable on the principal coordinate analysis (PCoA) plot ($p < 0.05$ except for Tibetan vs. Han from the same site, PERMANOVA based on JSD, Fig. 4b). The inter-site distance for samples belonging to the same ethnic groups was greater than the inter-ethnicity distance for samples from the same sites ($p < 0.0001$ for Uygur, Hui, and Mongolian), and correspondingly, the sampling site explained larger variance in the gut microbiota than the ethnic group (R^2 of PERMANOVA: Uygur, 7.70% vs 6.14%; Hui, 10.08% vs 3.81%; Mongolian, 9.33% vs 4.63%; Tibetan, 14.44% vs 4.88%) (Fig. 4b). These observations indicated that both geography and ethnicity could affect the gut microbiota, but the former is likely to have a stronger effect.

The genus-level microbiota profile of each ethnic group showed a distinct pattern in relative abundances, and we were able to distinguish different ethnic groups using the random forest model (AUC of the model, 0.88; Miao, 0.94; Uygur, 0.93; Bai, 0.93; Tibetan, 0.92; Naxi, 0.91; Hui, 0.82; Han 0.80; Mongolian, 0.80; Fig. 4c, Supplementary Fig. 4e). *Catenibacterium* contributed the most power to the classification, and the representative genera of enterotypes, *Bacteroides* (E2), *Escherichia* (E3), and *Prevotella* (E1), ranked 5th, 8th, and 9th of the contributing genera, respectively. Correspondingly, the enterotype composition differed among ethnic groups (Fig. 4d). Comparing to Han, Miao, Naxi, and Tibetan had higher proportions of E1; Miao, Naxi, Uygur, Hui, and Tibetan had lower proportions of E2; Miao had higher proportions while

Tibetan had lower proportions of E3; all ethnic minority groups except Naxi had higher proportions of E4 ($p < 0.05$, Fisher's exact test). We further applied DESeq2 models to detect ethnicity-specific genera by comparing one ethnic group to the rest, with adjustment for the confounding factor sampling site; for Uygur, Hui, Mongolian, and Tibetan, we also compared each of them to their accompanying Han samples from the same sampling sites. Differential genera ($p_{\text{adj}} < 0.05$) detected by both models included a lower level of *Clostridium XVIII* in Uygur, lower levels of *Prevotella*, *Fecalibacterium*, and *Alistipes*, whereas a higher level of *Romboutsia* in Hui, and higher levels of *Holdemanella* and *Enterococcus* whereas a lower level of *Escherichia* in Mongolian (Fig. 4c). Of note, the one-versus-rest comparison could be affected by the uneven sample size across ethnic groups.

Association between urbanization and the gut microbiota

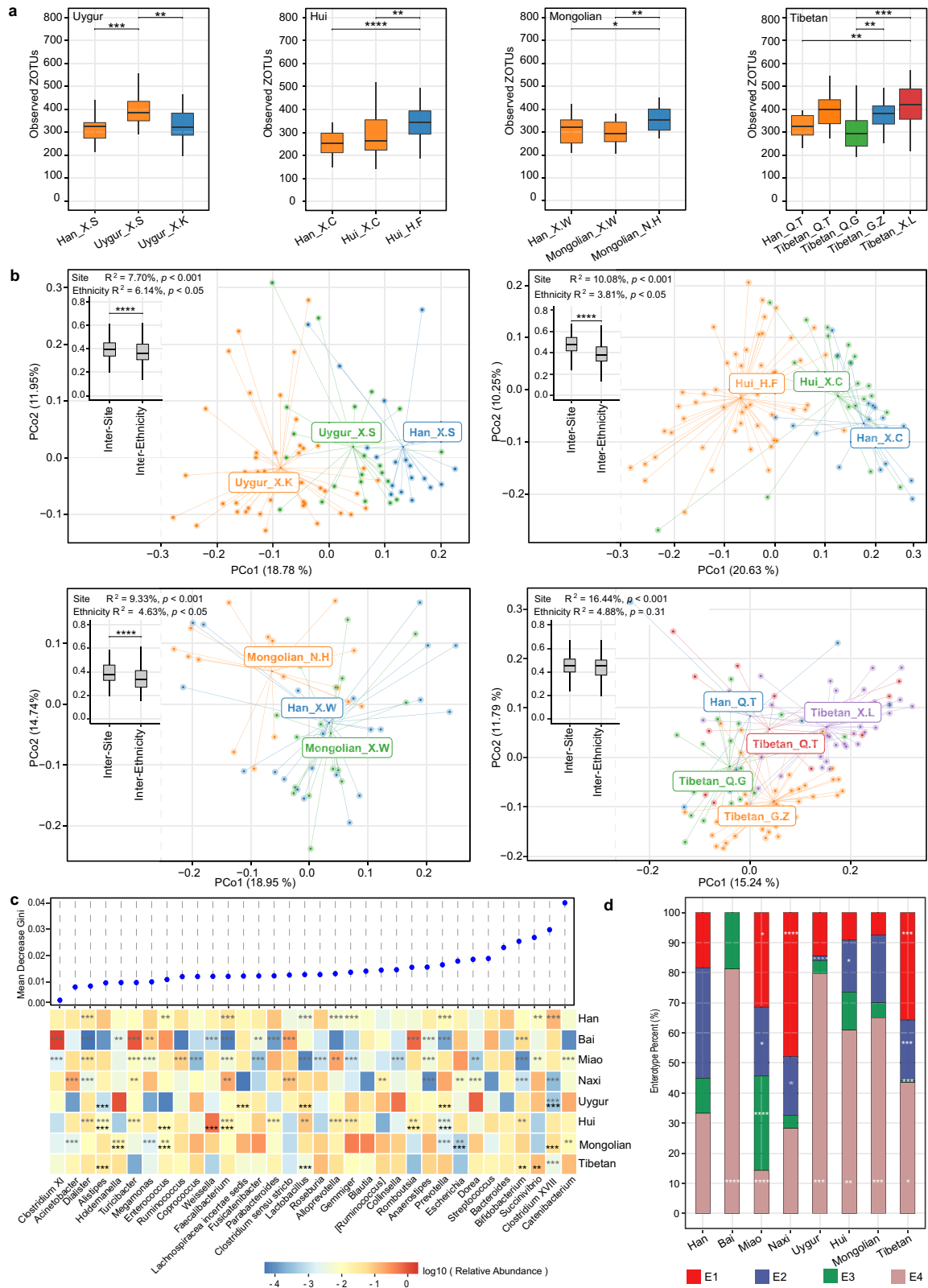
By comparing the gut microbiota of 1530 residents from 38 rural sites of 24 provinces and 637 residents from 22 urban sites of 18 provinces, we found that Faith's PD of rural residents was higher than that of urban residents, but Observed ZOTUs and Shannon index did not differ (Fig. 5a, Supplementary Fig. 5a–c). It suggested that urbanization might not affect the non-phylogenetic richness and evenness of the gut microbiota, but instead decrease its phylogenetic richness. Meanwhile, the overall gut microbiota composition differed between rural and urban residents (PERMANOVA based on JSD, $R^2 = 1.63$, $p < 0.01$). Of note, the intra-group microbiota dissimilarity evaluated by JSD was higher in the urban residents ($p < 0.001$, Fig. 5b).

The microbiota communities were compared using DESeq2 analysis to determine genera segregating urban and rural populations ($p_{\text{adj}} < 0.05$, Fig. 5c). Of the 33 differential genera, the top two most abundant (mean relative abundance $> 4.6\%$) ones were *Bacteroides* that were enriched in the urban population, and *Prevotella* that was enriched in the rural population. Besides, a few low-abundant genera (relative abundance $< 0.8\%$) showed more significant differences between the two groups ($\log_2FC > 1.5$, $p_{\text{adj}} < 1e-20$), including *Erysipelotrichaceae incertae sedis* and *Parasutterella* that were higher in the urban population, as well as *Alloprevotella* and *Catenibacterium* that were higher in the rural population.

Besides differences in the microbial composition mentioned above, we wonder if microbial interactions were also altered by urbanization, and thus constructed co-occurrence networks by SparCC for urban and rural populations separately (Fig. 5d). The rural population showed a larger network than the urban population (27 nodes with 36 edges vs 23 nodes with 23 edges), and random subsampling of the rural group to the equal number of samples in the urban group confirmed such a difference. Eighteen of the nodes were shared by the two networks, but only ten of the edges were shared, suggesting that correlations between the same microbial pairs were different between the two populations. Hub nodes also differed between the two networks. In the urban population network, two short-chain fatty acids (SCFA) producers, *Roseburia* and *Faecalibacterium*³⁴, as well as *Escherichia* connected with more edges, whereas the other three SCFA producers, *Blautia*, *Anaerostipes*, and *Dorea*^{34,35}, as well as *Clostridium XVIII* connected with more edges in the rural population network, suggesting different ecological assemblies supporting SCFA production in the gut ecosystem of the two populations.

DISCUSSION

We have conducted a study on Chinese gut microbiota with a large cohort covering by far the greatest diversity of the healthy population. We found that a series of factors belonging to five categories, i.e., geography, demography, diet, urbanization, and



sampling month, explained a substantial proportion of the gut microbiota variation, although the effect sizes of some factors (e.g., ethnicity) were likely under-estimated owing to the uneven number of samples in some subgroups. First, geographic factors showed the strongest signals, including sampling site,

geographical zone, altitude etc. Specifically, the gut microbiota of Han Chinese and ethnic minority groups from the same sampling sites was more alike than that of the same ethnic minority groups from different sampling sites, underscoring the importance of considering the geographic location in case-control

Fig. 4 Gut microbiota characteristics of different ethnic groups. **a** Observed ZOTUs of different ethnic groups per sampling site. The labels of x axis indicate ethnic groups followed by sampling sites. * $p_{\text{adj}} < 0.05$, ** $p_{\text{adj}} < 0.01$, *** $p_{\text{adj}} < 0.001$, Mann–Whitney test. **b** PCoA plot based on JSD. Inserted boxplots show the inter-site distance for samples belonging to the same ethnic groups, and inter-ethnicity distance for samples from the same sampling sites; **** $p < 0.0001$, Mann–Whitney test. Corresponding R^2 and p values from PERMANOVA tests were also shown. **c** Differences in the abundances of predominant genera between ethnicity groups. Genera were ordered by mean decrease in Gini from random forest models classifying ethnic groups. Genera with average relative abundance >1% and presence in >50% of the samples in at least one ethnic group were shown. Asterisks represent p_{adj} values from DESeq2 models; the gray ones indicate comparisons between one ethnic group and the rest groups (and adjusted for sampling sites), while the black ones indicate comparisons between each of the ethnic minority groups (Hui, Mongolian, Tibetan, and Uygur) and their accompanying Han samples from the same sampling sites; ** $p_{\text{adj}} < 0.01$, *** $p_{\text{adj}} < 0.001$. **d** Percentage of enterotypes in each ethnic group. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, Fisher's exact test comparing ethnicity minority groups to Han. In boxplots, the center line represents the median, box limits represent upper and lower quartiles and whiskers represent 1.5 \times interquartile range. **c–d** Han: $n = 1755$, Bai: $n = 16$, Miao: $n = 70$, Naxi: $n = 46$, Uygur: $n = 69$, Hui: $n = 87$, Mongolian: $n = 40$, Tibetan: $n = 154$. See also Figure S4.

studies. Although gut microbiota has been widely reported to vary across geography³⁶, it is hard to dissect the effect, as geography reflects a mixed effect of lifestyle, long-term diets etc. As to this cohort, we were not able to clarify patterns related to geographical zones, and the effect of altitude was likely linked to minority ethnic groups living in a plateau. Second, we focused on the ethnic group, which represents a highly diverse demographic character of the Chinese population. The Han Chinese and seven ethnic minority groups showed distinct gut microbiota profiles, with some of the variation being attributed to geography and a considerable part remaining significantly explained by the ethnic group. Larger cohorts and targeted design are required to understand the effect of covariates underlying ethnicity (e.g., genetics, custom) on the gut microbiota, and to what extent were these effects homogenized by, for example, co-residing with Han Chinese. Furthermore, our finding highlighted that urbanization was associated with decreased intra-individual diversity and increased inter-individual diversity of the gut microbiota. Previous studies based on one or two provinces of China have shown a similar pattern in the bacterial, fungal, and viral components^{15,37,38}, whereas our study based on 28 provinces further validated the vast effect of urbanization on the gut microbiota across China. In addition, although sampling month appeared to be a notable signal, we did not find any specific seasonal pattern in alpha or beta diversity. In contrast, the seasonal rhythm of gut microbiota and its CAZyme diversity has been shown in hunter-gatherers³⁹. We speculate that the lack of association between the gut microbiota and season in this cohort is because lifestyle, especially diet, is much less affected by season in the modernized population.

It is well acknowledged that diet alters gut microbial composition and metabolism, but the study on the long-term effect of staple food type on gut microbiome at the population level is still missing^{27,40}. The Chinese population mainly consumes two distinct types of staple food, wheat (products made from white flour of common wheat) and rice (boiled white rice). In this study, the wheat-consuming population was predicted to have a remarkably higher level of microbial genes encoding glucan 1,3-beta-glucosidase, the substrate of which only exists in wheat but not rice. This confirmed the validity of not only differences in the amount of wheat intake acquired by questionnaires, but also functional prediction based on 16S rRNA gene profiles. Further, the predicted increase in the microbial biosynthesis capacity of L-methionine and its major downstream product SAM in the Wheat group is of special interest, owing to the broad involvement of SAM in cognitive and metabolic health³². Notably, the alteration of microbial L-methionine biosynthesis as well as the archaeal conversion of SAM has been shown in an intervention study with gliadin (one of the protein fractions of wheat) in mice, which was based on urinary metabolome with ultra performance liquid chromatography-mass spectrometry⁴¹. Therefore, the possible effect of staple food on SAM biosynthesis may be narrowed down to the effect of the protein component of wheat, gluten,

which could trigger celiac disease in 0.06% of the Chinese population⁴². Thus, we speculate that gluten may modulate the health of gluten-tolerant individuals by regulating the gut microbiota, the validity, and mechanism of which warrant further investigation. In addition, since wheat is more popular in north China than in south China, the observed effect of staple food type on gut microbiota could be confounded by geographic locations.

We identified specific components of the microbiota that were significantly affected by the above factors. The three representative genera of enterotypes, *Bacteroides*, *Prevotella*, and *Escherichia* drove the diversification of the gut microbiota of the Chinese population (Fig. 6). Abundances of the three genera varied in subpopulations consuming different staple food, in different ethnic groups, and in urban vs rural residents. It is likely due to that dietary habits might be the shared covariate underlying these factors, and the close relation between *Bacteroides*/*Prevotella* and dietary habits especially fiber, protein, and animal fat has been widely shown¹⁹. Another remarkable genus is *Catenibacterium*, which was found to be more abundant in the population resided in rural areas, and the population consuming wheat as a staple food in this study. It was also the most differential genus among eight ethnic groups in this cohort. *Catenibacterium* was detected in 41% of the population with an average relative abundance of 0.6%. Limited studies on this genus have associated it with diet, but with conflicting results. A few studies showed its association with the Mediterranean diet and low risk of cardiovascular disease^{43,44}, whereas the others showed its association with a high-fat, high-sugar diet⁴⁵. These findings address the need for further investigation of *Catenibacterium* in relation to diet and human health. In addition, *Bifidobacterium* was associated with wheat intake, and it also contributed significantly to the differentiation of enterotypes and the eight ethnic groups. Of note, the overall abundance of *Bifidobacterium* in our cohort was considerably higher than that in the AGP cohort and other western cohorts^{17,46}, which may further strengthen its effect on health in specific subgroups of the Chinese population.

This study is limited by the absence of comprehensive physiological indices and detailed dietary information. Such information is essential in understanding factors responsible for or affected by the singular gut microbiota characteristics uncovered here, e.g., the enterotype *Escherichia* that was rarely reported in other cohorts, *Catenibacterium* that was enriched in specific subsets of the Chinese population, microbial SAM biosynthesis that possibly exerts an influence on human health. Nevertheless, the obtained profiles have constructed the Chinese gut microbiota baseline, providing information on the microbiota variation, covariates, and the effect size of covariates, which are essential for calculating sample size and statistical power for biomedical studies^{47,48}. Moreover, this study has raised attention to considering differences in the microbiota background and confounding factors, including those generally applicable or specific to the Chinese population, in microbiota researches and clinical translations.

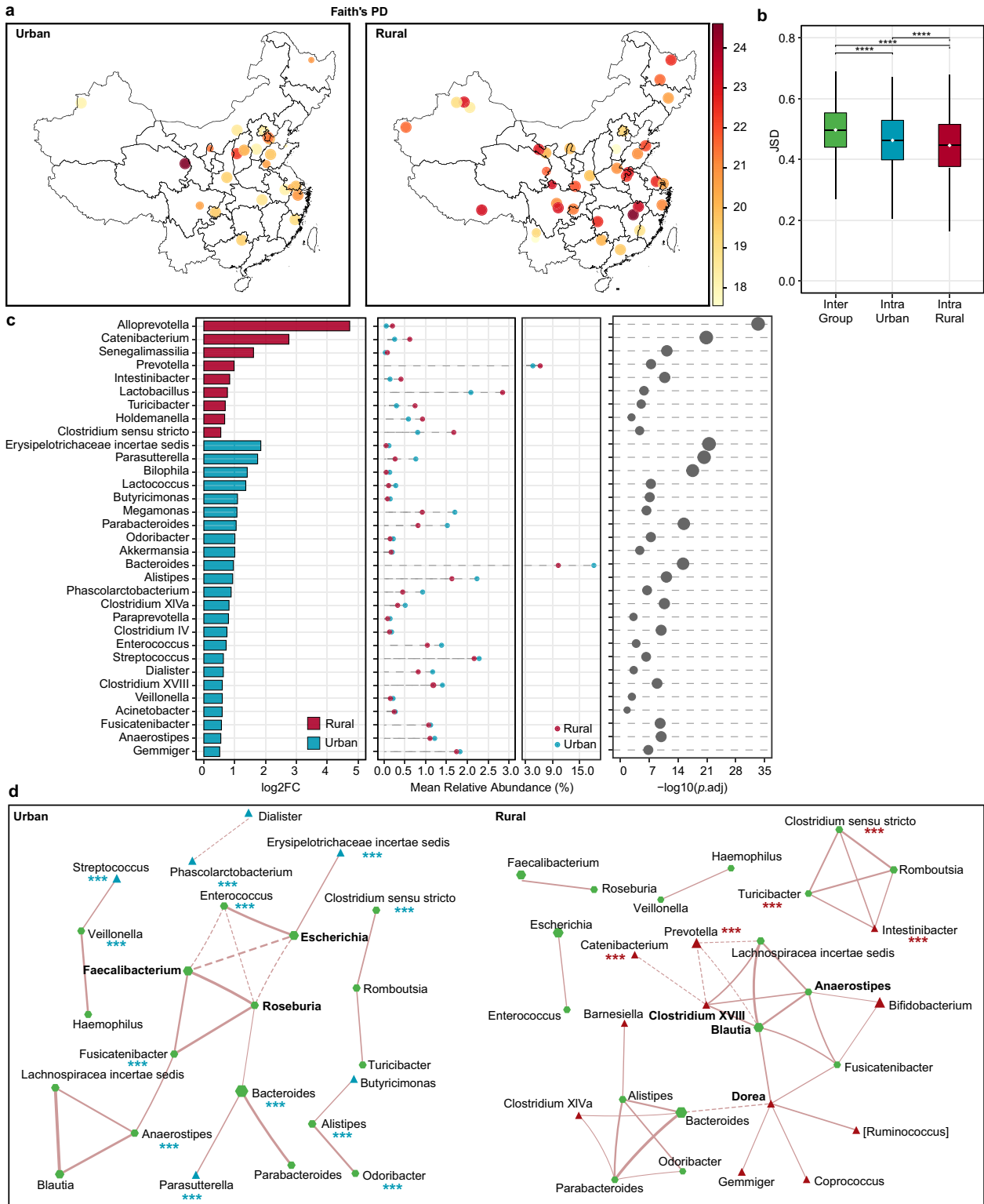


Fig. 5 Differences in the microbiota diversity, composition, and network between urban and rural populations. **a** Faith's PD. Each dot represents a sampling site; the color of the dot indicates the median value in each site; the diameter of the dot is proportional to the number of samples in each site and is fixed for sites with more than 15 samples. **b** Inter- and intra-group JSD. The intra-group distance was calculated per sampling site. **** $p < 0.0001$, Mann–Whitney test. The center line of the boxplot represents the median, box limits represent upper and lower quartiles and whiskers represent 1.5 \times interquartile range. **c** Differential genera were detected with the DESeq2 model. Genera with p . adj < 0.05 were shown. **d** Co-occurrence network of genera. SparCC correlations with $r > 0.35$ and $p < 0.01$ were shown. Each node represents a genus; the size of the node is proportional to the median relative abundance of the corresponding genus; the green, red, and blue colors represent the shared genus, rural-specific genus, and urban-specific genus, respectively. The solid and dotted lines represent positive and negative correlations, respectively, and the thickness of the edge is proportional to the r value. Red and blue asterisks indicate differential genera shown in **c**, **** p .adj < 0.001 . Urban: $n = 637$, rural: $n = 1530$. See also Figure S5.

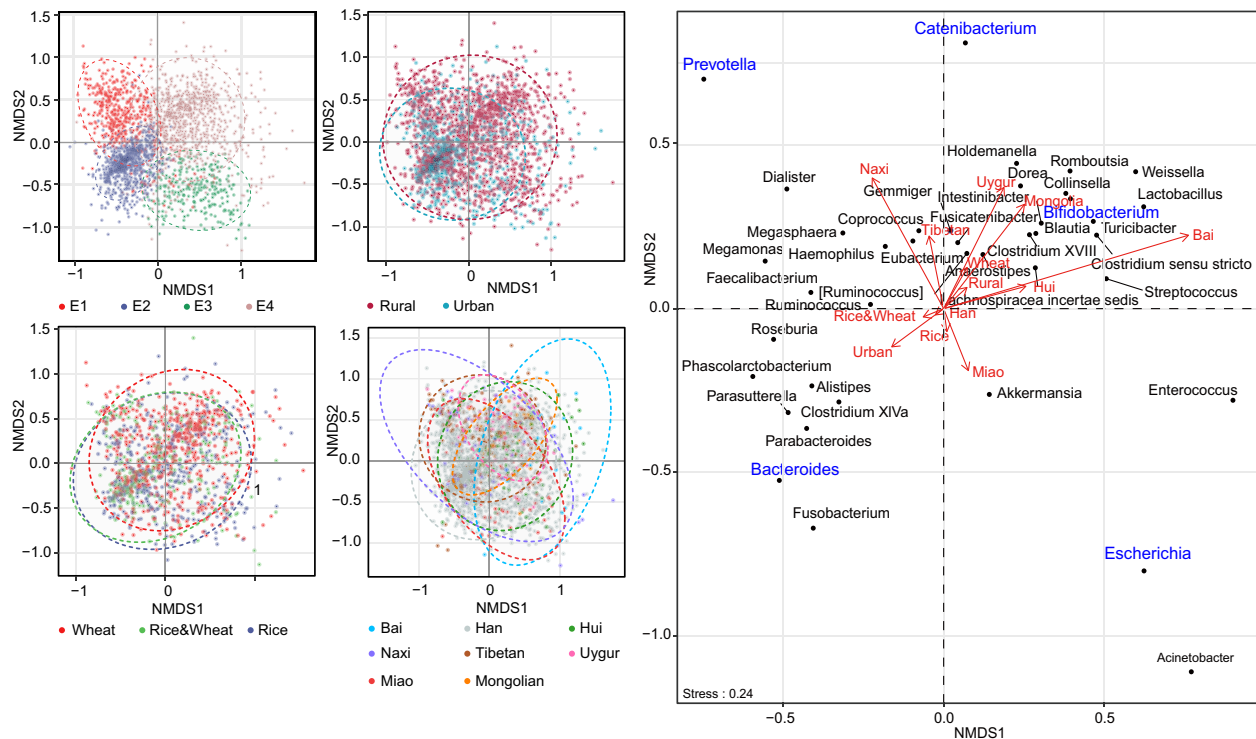


Fig. 6 Nonmetric multidimensional scaling plots showing clustering of Chinese gut microbiota by enterotype, urban/rural residence, staple food type, and ethnic group. The ordination was performed with Bray–Curtis distance based on the top 40 genera. Covariates were posteriorly projected onto the plot using envfit (vegan). Ellipses represent 95% confidence.

METHODS

The cohort and ethics

The 2678 Chinese participants had no self-reported gastrointestinal tract disorder or any other acute/chronic/recurrent medical conditions (referred to as “healthy”), and they had not taken any antibiotics for at least 3 months prior to participation. All recruited rural or pastoral residents lived a typical farming or pastoral lifestyle away from the metropolitan areas. Fecal samples were collected following a standardized procedure: the participants were informed of detailed instructions, collected samples by themselves, and stored samples in home freezers or iceboxes; samples were transported to the freezer at each sampling site within a day, and further to the research laboratory with cold-chain within 3 days; samples were then well homogenized, aliquoted, and stored at -80°C until further analyses. The metadata (Supplementary Data 1) was collected via questionnaire, including demographic information (age, gender, ethnicity, BMI, BMI z scores), dietary information (staple food type, intake of lactic acid bacteria (LAB)), and lifestyle (defecate frequency, sleep quality, alcohol intake, smoking). Specifically, the ethnic information was confirmed through participants’ Resident Identity cards. BMI z scores were calculated for individuals younger than 18 years old using the Zanthro function of STATA package v15.0, based on the World Health Organization Child Growth Standards 2007. Urbanization information (urban/rural/pastoral residence, population density, GDP) was collected from the National Bureau of Statistics of China and China Statistical Yearbook (2017 and 2018). Geographic information (sampling site, province, geographical zone, altitude, latitude, longitude) and sampling month were collected in the meantime. Geographic distances between sampling sites were represented by Vincenty distances, which were computed using the geosphere R package v1.5-10⁴⁹.

The study was approved by the Ethical Committee of Jiangnan University. Written informed consents were obtained from all participants or their legal representatives for minors.

16S rRNA gene sequencing

Microbial DNA was extracted from feces using the MP FastDNA Spin Kit for Feces (MP Biomedicals, Santa Ana, CA, USA) following the manufacturer’s instructions. The V3–4 region of 16S rRNA gene was amplified by the primers 314 F (CCTAYGGRRBGCASCAG) and 806 R

(GGACTACNNGGGTATCTAAT) joined with a seven-base-pair barcode. The PCR product was purified by the QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany), and sequenced on the Illumina Miseq platform with the Miseq Reagent Kit V3 (Illumina, San Diego, CA, USA, PE300 mode).

Microbiota data analysis

Paired-end sequencing reads were merged using USEARCH v11.0.667⁵⁰. The reads were de-multiplexed, and barcode and primer sequences were removed with Cutadapt v2.11⁵¹. All sequences were subjected to quality filtering with a cutoff of maxee 1.0 and de-replicated, and unique sequences with more than seven replicates were clustered to ZOTUs using USEARCH. Taxonomy of ZOTU representative sequences was assigned using the SINTAX algorithm of USEARCH and the adjusted Ribosomal Database Project (RDP) training set v16 (https://github.com/Li-Zhang/rdp_16s_v16_sp_ManualAdjustment) with a cutoff of 0.8. The phylogenetic tree was constructed by inserting ZOTU representative sequences into the 99% Greengenes 13_8 reference tree using the SEPP algorithm⁵² with QIIME2 v2018.10⁵³. The functional potential was predicted based on ZOTUs using PICRUSt²⁹, generating EC number and MetaCyc pathway abundances.

For analyzing the microbiota alpha and beta diversity, the ZOTU table was rarefied to 10,000 reads, and observed ZOTUs, Faith’s PD, Shannon index, Bray–Curtis distance, and unweighted UniFrac distance were estimated using QIIME2, whereas JSD was estimated using the phyloseq R package v1.32.0⁵⁴. The microbiota variation explained by environmental variables was evaluated with envfit function of the vegan R package v2.5–6²¹, and the differential clustering of microbial communities was assessed using PERMANOVA with adonis function of vegan. The Correlation between geographic distances and microbial JSD was estimated by Mantel statistic based on Spearman’s rank correlation, using mantel function of vegan.

For the rest of the analyses, total sum scaling was applied to normalize the microbiome data unless otherwise specified. Enterotype analysis was done based on the method described by Arumugam et al.¹⁴. Samples were clustered with the pam function of the cluster R package v2.1.0⁵⁵. The optimal number of clusters was four according to the Calinski–Harabasz Index, giving an average silhouette coefficient of 0.14. The clustering was visualized on PCoA using the ade4 R package v1.7-15⁵⁵. To identify the

driving genera of each enterotype, random forest analysis with ten-time fivefold cross-validation was performed using the randomForest R package v4.6-14⁵⁶. The representative genus *Escherichia/Shigella* was referred to as *Escherichia*, considering that all subjects had no symptoms of *Shigella* infection at the time of sampling.

Random forest models with ten times fivefold cross-validation was applied to identify genera distinguishing the eight ethnic groups using the scikit-learn Python package v0.23.1⁵⁷. To even out the number of individuals in each ethnic group, the Han Chinese were randomly downsampled to 6% of the full data set 1000 times. Accordingly, the Mean Decrease in Gini, AUC, sensitivity, specificity, and precision were calculated as the average value from 1000 random forest models.

To examine the microbiota co-occurrence network, correlation analysis was performed on genera (data rarefied to 10,000 reads) using SparCC⁵⁸, and visualized using Cytoscape v3.5.1⁵⁹. Differential genera were detected using the DESeq2 R package v1.29.14⁶⁰, with adjustment for age and gender unless otherwise specified. Genera with mean relative abundance >0.1% and presence in >50% of the samples in at least one group were used for the above analyses unless otherwise specified.

Statistical analysis

Two-tailed Mann–Whitney test and Fisher’s exact test were used to compare continuous variables and categorical variables, respectively. Multiple comparisons were corrected using the Benjamini–Hochberg false discovery rate algorithm⁶¹ with a significance level of 0.05 (*p*.adj value).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The microbial DNA sequences encoding 16 S rRNA V3–4 region reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center⁶², Beijing Institute of Genomics (China National Center for Bioinformatics), Chinese Academy of Sciences, under accession number CRA003616 that are publicly accessible at <https://bigd.big.ac.cn/gsa>.

Received: 10 March 2021; Accepted: 11 August 2021;

Published online: 06 September 2021

REFERENCES

- Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. *Genome Med.* **8**, 51 (2016).
- Rajilić-Stojanović, M., Heilig, H. G., Tims, S., Zoetendal, E. G. & de Vos, W. M. Long-term monitoring of the human intestinal microbiota composition. *Environ. Microbiol.* **15**, 1146–1159 (2012).
- Faith, J. J. et al. The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).
- 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. *Nat. Commun.* **8**, 1784 (2017).
- Wong, S. H. & Yu, J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 690–704 (2019).
- He, Y. et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).
- Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
- Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
- Wu, G. D. et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108 (2011).
- Ayeni, F. A. et al. Infant and adult gut microbiome and metabolome in rural bassa and urban settlers from Nigeria. *Cell Rep.* **23**, 3056–3067 (2018).
- Deschasaux, M. et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nat. Med.* **24**, 1526–1531 (2018).
- Cadwell, K., Brooks, A. W., Priya, S., Blekhan, R. & Bordenstein, S. R. Gut microbiota diversity across ethnicities in the United States. *PLoS Biol.* **16**, e2006842 (2018).
- Brewster, R. et al. Surveying gut microbiome research in africans: toward improved diversity and representation. *Trends Microbiol.* **27**, 824–835 (2019).
- Zhang, J. et al. A phylo-functional core of gut microbiota in healthy young Chinese cohorts across lifestyles, geography and ethnicities. *ISME J.* **9**, 1979–1990 (2015).
- Winglee, K. et al. Recent urbanization in China is correlated with a Westernized microbiome encoding increased virulence and antibiotic resistance genes. *Microbiome* **5**, 121 (2017).
- Derrien, M., Alvarez, A. S. & de Vos, W. M. The gut microbiota in the first decade of life. *Trends Microbiol.* **27**, 997–1010 (2019).
- Xiao, L. et al. A catalog of the mouse gut metagenome. *Nat. Biotechnol.* **33**, 1103–1108 (2015).
- Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
- Costea, P. I. et al. Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2018).
- Liang, C. et al. Diversity and enterotype in gut bacterial community of adults in Taiwan. *BMC Genomics* **18**, 932 (2017).
- Oksanen, J. et al. Package ‘vegan’. *Community Ecol. package, version 2*, 1–295 (2013).
- Zhang, Y., Wang, Y. & Niu, H. Effects of temperature, precipitation and carbon dioxide concentrations on the requirements for crop irrigation water in China under future climate scenarios. *Sci. Total Environ.* **656**, 373–387 (2019).
- Collado, M. C., Donat, E., Ribes-Koninckx, C., Calabuig, M. & Sanz, Y. Imbalances in faecal and duodenal Bifidobacterium species composition in active and non-active coeliac disease. *BMC Microbiol.* **8**, 232 (2008).
- Di Cagno, R. et al. Different fecal microbiotas and volatile organic compounds in treated and untreated children with celiac disease. *Appl Environ. Microbiol.* **75**, 3963–3971 (2009).
- Staudacher, H. M. et al. A diet low in FODMAPs reduces symptoms in patients with irritable bowel syndrome and a probiotic restores bifidobacterium species: a randomized controlled trial. *Gastroenterology* **153**, 936–947 (2017).
- Hansen, L. B. S. et al. A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. *Nat. Commun.* **9**, 4630 (2018).
- Li, J. et al. Carbohydrate staple food modulates gut microbiota of mongolians in China. *Front. Microbiol.* **8**, 484 (2017).
- Mano, F. et al. The effect of white rice and white bread as staple foods on gut microbiota and host metabolism. *Nutrients* **10**, 1323 (2018).
- Douglas, G. M. et al. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688 (2020).
- Kaur, B., Ranawana, V. & Henry, J. The glycemic index of rice and rice products: a review, and table of GI values. *Crit. Rev. Food Sci. Nutr.* **56**, 215–236 (2016).
- Shewry, P. R., Pellny, T. K. & Lovegrove, A. Is modern wheat bad for health? *Nat. Plants* **2**, 16097 (2016).
- Anstee, Q. M. & Day, C. P. S-adenosylmethionine (SAME) therapy in liver disease: a review of current evidence and clinical utility. *J. Hepatol.* **57**, 1097–1109 (2012).
- Sharma, A. et al. S-Adenosylmethionine (SAME) for neuropsychiatric disorders: a clinician-oriented review of research. *J. Clin. Psychiatry* **78**, e656–e667 (2017).
- Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* **165**, 1332–1345 (2016).
- Taras, D., Simmering, R., Collins, M. D., Lawson, P. A. & Blaut, M. Reclassification of *Eubacterium formicigenans* Holdeman and Moore 1974 as *Dorea formicigenans* gen. nov., comb. nov., and description of *Dorea longicatena* sp. nov., isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **52**, 423–428 (2002).
- Gupta, V. K., Paul, S. & Dutta, C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front. Microbiol.* **8**, 1162 (2017).
- Sun, Y. et al. Population-level configurations of gut mycobiome across six ethnicities in urban and rural China. *Gastroenterology* **28**, 741–751.e4 (2020).
- Zuo, T. et al. Human-gut-DNA virome variations across geography, ethnicity, and urbanization. *Cell Host Microbe* **160**, 272–286.e11 (2020).
- Smits, S. A. et al. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802–806 (2017).
- Bremme, K., Eneroth, P., Carsjö, B. M. & Nilsson, B. Blood pressure regulation in third-trimester pregnant women receiving tocolytic terbutaline infusion. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **23**, 53–60 (1986).
- Zhang, L. et al. Effects of gliadin consumption on the intestinal microbiota and metabolic homeostasis in mice fed a high-fat diet. *Sci. Rep.* **7**, 44613 (2017).
- Yuan, J. et al. Prevalence of celiac disease autoimmunity among adolescents and young adults in China. *Clin. Gastroenterol. Hepatol.* **15**, 1572–1579.e1571 (2017).
- García-Manzana, I., Selma-Royo, M., Alcantara, C. & Collado, M. C. Shifts on gut microbiota associated to mediterranean diet adherence and specific dietary intakes on general adult population. *Front. Microbiol.* **9**, 890 (2018).

44. Shankar, V. et al. Differences in gut metabolites and microbial composition and functions between Egyptian and U.S. children are consistent with their diets. *mSystems* **2**, e00169–16 (2017).
45. Turnbaugh, P. J. et al. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1**, 6ra14 (2009).
46. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
47. Mattiello, F. et al. A web application for sample size and power calculation in case-control microbiome studies. *Bioinformatics* **32**, 2038–2040 (2016).
48. Kelly, B. J. et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics* **31**, 2461–2468 (2015).
49. Hijmans, R. J. in Introduction to the “geosphere” package (Version 1.5-10) (2019).
50. Edgar, R. C. & Flyvbjerg, H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31**, 3476–3482 (2015).
51. Kechin, A., Boyarskikh, U., Kel, A. & Filipenko, M. cutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing. *J. Comput. Biol.* **24**, 1138–1143 (2017).
52. Janssen, S. et al. Phylogenetic placement of exact amplicon sequences improves associations with clinical information. *mSystems* **3**, e00021–18 (2018).
53. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
54. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).
55. Brock, G., Pihur, V., Datta, S. & Datta, S. cValid, an R package for cluster validation. *Journal of Statistical Software* (Brock et al., March 2008) (2011).
56. RColorBrewer, S. & Liaw, M. A. Package ‘randomForest’. University of California, USA (2018).
57. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
58. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
59. Cline, M. S. et al. Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366 (2007).
60. Love, M., Anders, S. & Huber, W. Differential analysis of count data—the DESeq2 package. *Genome Biol.* **15**, 10.1186 (2014).
61. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
62. Partners, N. G. D. C. Ma Database resources of the national genomics data center in 2020. *Nucleic Acids Res.* **48**, D24–d33 (2020).

ACKNOWLEDGEMENTS

This work was funded by the National Natural Science Foundation of China (Grant nos. 31771953, 31820103010), the Fundamental Research Funds for the Central Universities (JUSRP51903B), 111 project (BP0719028), the Collaborative Innovation Center of Food Safety and Quality Control in Jiangsu Province, and the National First-class Discipline Program of Food Science and Technology (JUFSTR20180102). The authors thank numerous professors and Ph.D. students

from the Research Center of Food Biotechnology at the School of Food Science & Technology of Jiangnan University for their help with questionnaire surveys and fecal sample collection. In addition, Linfei Zhou, Qingsong Zhang, and Qi Liu are acknowledged for their help with figure preparation; Jiaxin Zhong, Leisheng Shi, Mingluo Pan, Miaoqi Xie, Yutao Chen, and Tong Wu are acknowledged for their help with bioinformatics. Qinying He and Qundi Feng are acknowledged for their help with BMI z score calculation.

AUTHOR CONTRIBUTIONS

W.C., W.L., H.Z., J.Z., and Q.Z. conceived and designed the study. W.L. and J.L. managed the cohort, sample collection, and data acquisition. J.L., L.Z., M.L., and W.L. analyzed and interpreted the data. L.Z., M.L., and J.L. prepared the manuscript with the contribution from Y.-K.L. All authors approved the final manuscript. J.L. and L.Z. contributed equally.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41522-021-00245-0>.

Correspondence and requests for materials should be addressed to W.L., M.L. or W.C.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021