LETTERS
https://doi.org/10.1038/s41591-018-0164-x

nature medicine

Corrected: Author Correction

Regional variation limits applications of healthy gut microbiome reference ranges and disease models

Yan He[1,16], Wei Wu[2,3,16], Hui-Min Zheng[1,2,16], Pan Li[1,2,16], Daniel McDonald[4], Hua-Fang Sheng[1], Mu-Xuan Chen[1], Zi-Hui Chen[3], Gui-Yuan Ji[3], Zhong-Dai-Xi Zheng[2], Prabhakar Mujagond[5], Xiao-Jiao Chen[1], Zu-Hua Rong[1,2], Peng Chen[6], Li-Yi Lyu[7], Xian Wang[7], Chong-Bin Wu[7], Nan Yu[1], Yan-Jun Xu[8], Jia Yin[9], Jeroen Raes[10,11,12], Rob Knight[4,13,14], Wen-Jun Ma[3]* and Hong-Wei Zhou[1,2,15]*

Dysbiosis, departure of the gut microbiome from a healthy state, has been suggested to be a powerful biomarker of disease incidence and progression[1–3]. Diagnostic applications have been proposed for inflammatory bowel disease diagnosis and prognosis[4], colorectal cancer prescreening[5] and therapeutic choices in melanoma[6]. Noninvasive sampling could facilitate large-scale public health applications, including early diagnosis and risk assessment in metabolic[7] and cardiovascular diseases[8]. To understand the generalizability of microbiota-based diagnostic models of metabolic disease, we characterized the gut microbiota of 7,009 individuals from 14 districts within 1 province in China. Among phenotypes, host location showed the strongest associations with microbiota variations. Microbiota-based metabolic disease models developed in one location failed when used elsewhere, suggesting that such models cannot be extrapolated. Interpolated models performed much better, especially in diseases with obvious microbiota-related characteristics. Interpolation efficiency decreased as geographic scale increased, indicating a need to build localized baseline and disease models to predict metabolic risks.

Commercial microbiota tests are available that identify dysbiosis by comparing the relative abundances of gut microbial taxa between customers and a small cohort of healthy individuals. However, dysbiosis patterns vary among studies[9–11] and are potentially affected by geographical location. Previous studies have compared western and nonindustrialized populations and found substantial variability in gut microbiota characteristics among these groups[12,13]. However, the generalizability of healthy baseline and disease model data among different locations has been investigated in only a few meta-analyses[14] and small-scale studies[15]. Additional major challenges are that: (i) microbiota sequencing results vary more among technical protocols than among populations[16]; (ii) diagnostic signals are affected by a wide range of confounders, including drugs[17], and (iii) human gut microbiota are highly variable, and large sample sizes are therefore needed to draw reliable conclusions. Hence, meta-analyses and small-scale studies, although enlightening, do not necessarily generalize to new populations[14].

Testing the generalizability of diagnoses based on microbiota differences between healthy and disease states requires data from studies performed with a regionalized study design, extensive sampling and standardized experimental protocols. We applied these principles in Guangdong province (GGMP, Guangdong Gut Microbiome Project), the most developed area in China (gross domestic product for 2017, ¥8.99 trillion ($1.33 trillion); comparable to South Korea). This province spans coastal and mountain areas and has an area of 179,800 km^2 and a population of 108 million people. We randomly selected 14 districts in this province, including both major cities and less-developed areas (Fig. 1a). To increase representativeness within each district, we first randomly selected three neighborhoods per district, then two communities per neighborhood, and finally 45 households per community. Volunteers ≥18 years old who were willing to participate were enrolled after providing informed consent. We used identical sample collection protocols across all participants and frozen transport for sample transfer. We assessed gut microbiota by sequencing the 16S rRNA gene variable 4 (V4) region. Using quantitative insights into microbial ecology (QIIME)[18], we analyzed 7,009 microbiota data samples together with 72 host parameters[18] (Supplementary Fig. 1 and Supplementary Table 1). A high level of Proteobacteria was observed in our population (Fig. 1a), which is unlikely due to improper sample transportation and DNA extraction, according to our validation (Methods and Supplementary Fig. 2).

We used Adonis to explore associations between variation in gut microbiota and host characteristics. Given a false discovery rate (FDR) of 5%, 51 parameters were significantly associated with gut microbial variations derived from between-sample weighted

[1]Division of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, China. [2]Department of Environmental Health, School of Public Health, Southern Medical University, Guangzhou, China. [3]Guangdong Provincial Institute of Public Health, Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China. [4]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [5]Gut Infection and Inflammation Biology Lab, UNESCO-Regional Center for Biotechnology, NCR Biotech Science Cluster, Faridabad, India. [6]Department of Pathophysiology, Southern Medical University, Guangzhou, China. [7]Shenzhen Fun-Poo Biotech Co., Ltd., Shenzhen, China. [8]Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China. [9]Department of Neurology, NanFang Hospital, Southern Medical University, Guangzhou, China. [10]Department of Microbiology and Immunology, KU Leuven–University of Leuven, Leuven, Belgium. [11]VIB, Center for the Biology of Disease, Leuven, Belgium. [12]Vrije Universiteit Brussel, Faculty of Sciences and Bioengineering Sciences, Microbiology Unit, Brussels, Belgium. [13]Department of Computer Science & Engineering, Jacobs School of Engineering, University of California San Diego, La Jolla, CA, USA. [14]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA. [15]State Key Laboratory of Organ Failure Research, Southern Medical University, Guangzhou, China. [16]These authors contributed equally to this work: Yan He, Wei Wu, Hui-Min Zheng, Pan Li. *e-mail: mawj@gdiph.org.cn; biodegradation@gmail.com

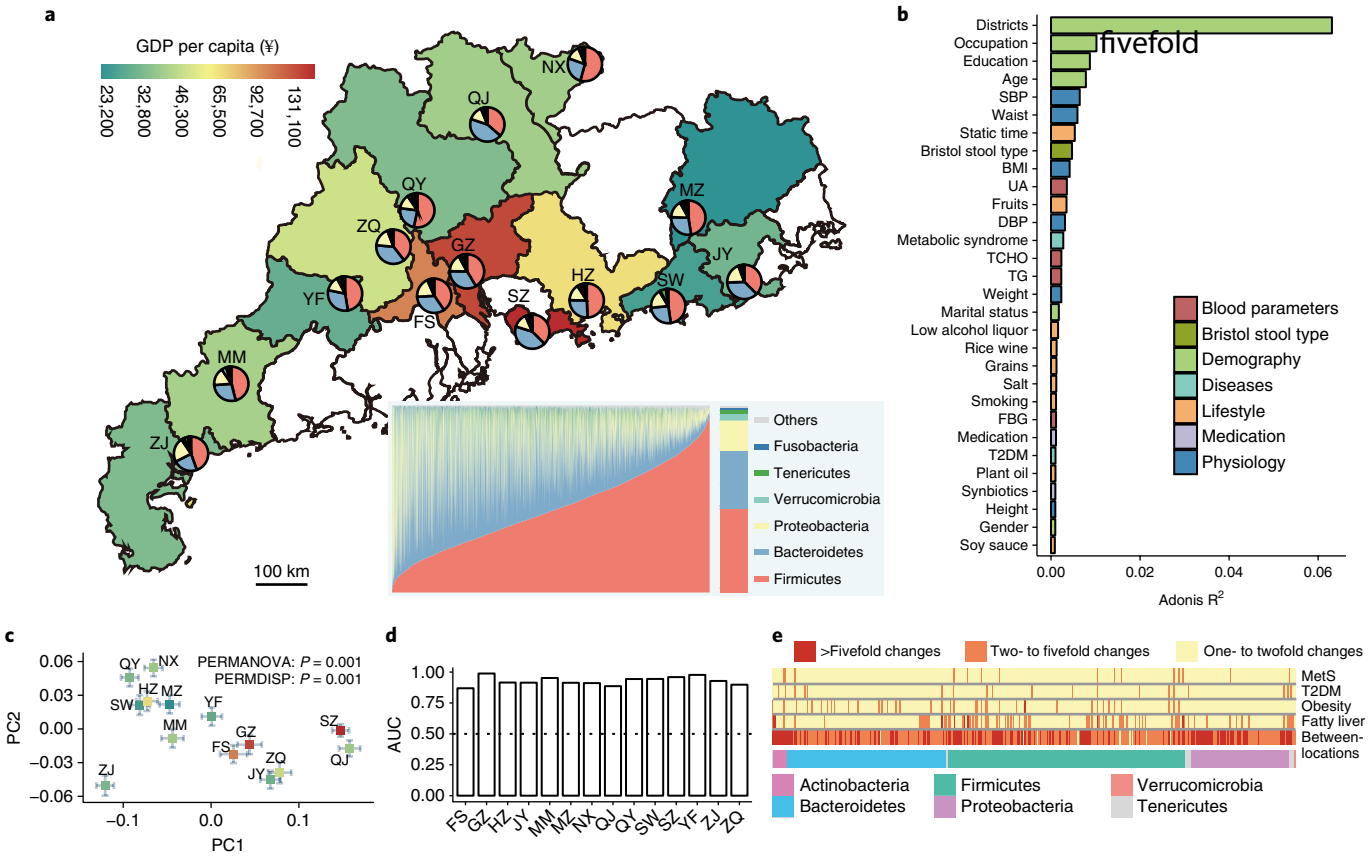

Fig. 1 | Overview of sampling regions and of regional variation in gut microbiota. a, Overview of sampling regions showing the 14 districts selected in the Guangdong province of southern China. Pie charts show the phylum-level microbiota found in each region. Background colors show local GDP per capita (¥). Variation in phyla among individuals is also shown plotted according to the abundance of Firmicutes. **b**, Bar plot illustrating the top 30 host factors found to be significantly associated with gut microbial variations. The variations were derived from between-sample weighted UniFrac distances. The bars are colored according to metadata categories. Size effect and statistical significance were calculated by PERMANOVA (Adonis). FDR was controlled at 5%. DBP, diastolic blood pressure; TCHO, total cholesterol; FBG, fasting blood glucose. **c**, Principal coordinate (PC) analysis showing the centroid of each district (the number of individuals in each district is $n = 500 \pm 71$ (mean ± s.d.)) as a mean of PC values and standard error (mean ± s.e.). Differences in beta-diversity and beta-dispersion among locations were tested by PERMANOVA (Adonis) and PERMDISP, respectively. **d**, Barplot of the area under ROC for each district. **e**, Heatmap showing maximal fold change differences in OTUs among different locations and metabolic disorders. Fold changes were grouped as one to two, two to five and more than five.

UniFrac distances (Fig. 1b, Supplementary Fig. 3 for unweighted UniFrac and Bray–Curtis distances and Supplementary Tables 2–4 for details). Age, Bristol stool type, body mass index (BMI), systolic blood pressure (SBP), triglyceride (TG) level and uric acid (UA) level were among the strongest explanatory factors, consistent with reports in Western populations[17,19]. However, host location exerted the strongest effect (Fig. 1b), far exceeding the effects of other host phenotypes. Four additional methods, including analysis of similarities (ANOSIM), multiresponse permutation procedures (MRPP), distance-based redundancy analysis (db-RDA) and envfit validated the robustness of the geographical effect (Supplementary Fig. 4). The ordination method showed that district centroids were scattered (Fig. 1c), indicating that the geographical effect was not caused by variations in beta-dispersion[20] or by single outliers. The effect of socioeconomic status, including occupation and education, varied according to location (Supplementary Fig. 5). We then compared the relative abundances of gut operational taxonomic units (OTUs) among the 14 districts using the Kruskal–Wallis test. Given a FDR of 5%, 794 out of 950 OTUs significantly discriminated the 14 districts and accounted for 99.1% of the sequences (Supplementary Table 5). We were further able to use a random forest model to identify a subject's location solely on the basis of fecal microbiota data (area under curve (AUC), 86.9–98.8%; Fig. 1d). Our nested study design enabled us to determine that the geographical effect affected 13 out of 14 districts and 25 out of 42 neighborhoods, indicating that in our study population, this effect began to emerge in comparisons between communities residing within the same neighborhood (Supplementary Fig. 6).

A common approach to clinical test development is establishing a normal range for healthy subjects. However, applying this strategy beyond the original reference population requires a consistent range. In our population, 2,008 individuals were healthy (no reported diseases, fasting blood glucose (FBG) < 6.1, BMI < 24, and no antibiotic use within 1 month of donating a fecal sample), and geographic location also dominated the variation observed among these 'healthy' gut microbiota profiles (Supplementary Fig. 7 and Supplementary Tables 6–8). Geographic variation effect sizes exceeded those of metabolic diseases, including type 2 diabetes (T2DM), metabolic syndrome (MetS), obesity and fatty liver. Given a FDR of 5%, 374 OTUs were significantly associated with at least one metabolic disease. Unfortunately, these OTUs were also significantly different among different locations (Supplementary Fig. 8 and Supplementary Tables 9–12). In our population, these biomarkers changed one- to twofold between healthy individuals and those

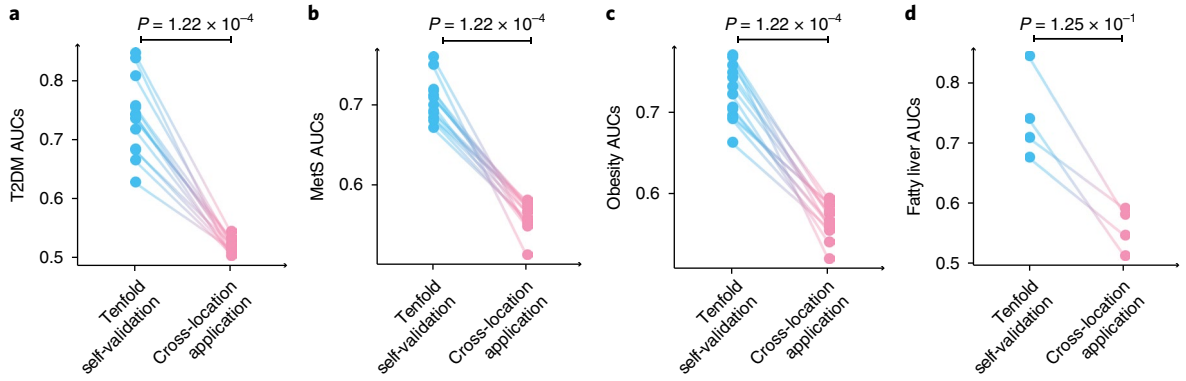

Fig. 2 | Evaluating cross-applicability of gut microbiota–based disease models among locations. a–d, Machine learning models were trained for T2DM (**a**), MetS (**b**), obesity (**c**) and fatty liver (**d**) in each of the 14 districts ($n=14$) and cross-applied to other districts ($n=14$). The fatty liver models were based on only four locations ($n=4$ for the self-validation and $n=4$ for the cross application) because of its low prevalence in other districts. Dotplots show the AUCs of tenfold cross-validations in the training site connected to the AUCs achieved when the same model was applied to other sites. Differences between the two groups for **a–d** were compared using the two-sided paired Wilcoxon rank sum test.

with metabolic disease, but 40.1% of them varied more than fivefold among locations (Fig. 1e). Therefore, healthy reference baselines for gut microbiota can be highly dependent on location, and it is risky to extrapolate from a single reference population.

This problem cannot be solved by integrating multiple microbiome biomarkers. We used microbiota profiles to construct accurate random forests models for T2DM within each district (AUC 0.63–0.85, mean 0.74; Fig. 2a). Applying a disease model trained in one location to another location dramatically decreased AUCs to ~0.5, comparable to random guessing (Fig. 2a). This inconsistency was not caused by metformin use[11] because participants taking medication were excluded from the model. Moreover, a prior study was unable to extrapolate for T2DM even after disentangling the effect of metformin[11], consistent with our findings. Applying the same approach to MetS (Fig. 2b), obesity (Fig. 2c) and fatty liver (Fig. 2d) yielded similar results. Consequently, our demonstration showed that building gut microbiota–based metabolic disease models in one location and applying them to others resulted in a significantly higher number of incorrect healthy status predictions compared to that when applying to the training set (Supplementary Figs. 9–12). These results indicate that diagnostic models of metabolic diseases trained in one location should not be applied in another location because of between-location variation in residents' gut microbiota.

Although extrapolations between districts did not yield usable predictions, interpolating the models from a larger geographic region might work better. Models trained at the province level had interpolated AUCs of 0.56 for T2DM, 0.64 for MetS, 0.66 for obesity and 0.74 for fatty liver, indicating a gradient corresponding to the fold-changes in OTUs induced by the four diseases (Fig. 1e and Supplementary Fig. 13). These results show that consistent signals, if found, could be extracted by the model for interpolation. A difficulty gradient was implied by this analysis, in that interpolating within a smaller scale yields much better results than extrapolating, and interpolating within a larger scale yields results of intermediate quality (Fig. 3a). Our nested design enabled us to examine this principle on different geographical levels with MetS models, and all levels conformed to this principle (Fig. 3b and Supplementary Fig. 14). It was previously reported that microbiota modeling of inflammatory bowel disease (IBD)[21] and colorectal cancer[22] could be extrapolated between different populations by meta-analysis, possibly because their size effect on changes in gut microbiota was larger than those of metabolic disorders[23]. These two diseases are not illustrated here because their sample sizes were too small in our population to yield reliable results. We therefore propose that future

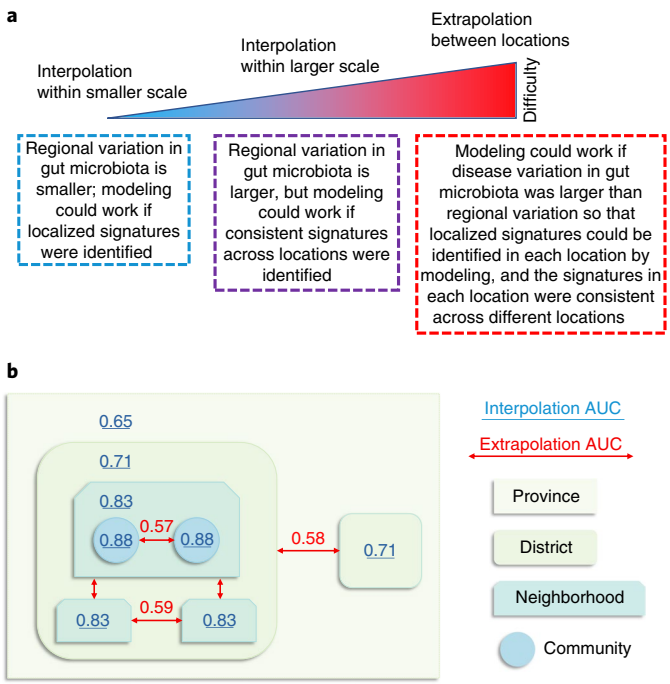

Fig. 3 | Illustration of the difficulty gradient used to interpolate and extrapolate the MetS model. a, Principles involved in the difficulty gradient of model show that it is much easier to interpolate within a smaller scale than to extrapolate between them, whereas interpolating within a larger scale lies between the two. The conditions that enable a workable interpolation and extrapolation are also described. **b**, In our nested study design, we selected 14 districts from the Guangdong province, 3 neighborhoods in each selected district and 2 communities in each neighborhood. Here, we illustrate the median AUC values for the interpolations and extrapolations for MetS in each geographic level. Details are shown in Supplementary Fig. 14.

studies should carefully evaluate the general applicability of gastrointestinal disease models on a case-by-case basis.

Collectively, these data show that geography exerts a strong effect on human gut microbiota variations, and this partly explains the inconsistent dysbiosis patterns reported in other small-scale studies[9–11]. Although it is difficult to extrapolate metabolic diseases

models between locations, interpolating these data from a training set collected on a larger geographic scale might work. Results will depend on identifying consistent signatures, which becomes more difficult as the geographic scale becomes larger. The geographical effect should be carefully considered in case–control studies because it could emerge on a small geographical scale and bias disease signals. One interesting question suggested by this geographical effect is: are regional gut microbiota characteristics are a link between geography and disease epidemiology? Future studies that explore relationships among human gut microbiota, economic development, lifestyle changes and disease epidemiology at the regional level will provide enlightening data for public health science. Notably, a strong geographical effect other than a small-scale distance decay effect was not observed in the American Gut Project (AGP)[24] but was observed in an Ecuadorian population[25], indicating that whether this effect is a universal phenomenon or not remains controversial. However, we note that the AGP was not designed with the type of stratified sampling that optimizes statistical power to address the questions asked in this study. Whether the geographical effect we report is driven by host-specific factors or introduced by other ecological processes, such as dispersal, drift, local diversification or host interactions with environmental microbiota, requires further investigation[26]. The present population was unique in its highly homogeneous ancestry (all participants were Chinese, and 99% were Han), which therefore allows us to isolate the effect of geography without ethnicity as a confounding factor[27] but conversely does not examine how ethnicity may confound geographical effects. This question should be examined in regions with more diverse genetic backgrounds, ideally where the same ethnic groups are found together in multiple geographic locations.

In conclusion, our data reinforce the need to use consistent sampling protocols to build localized reference baselines for gut microbiota. The applicability of disease models in new populations must be explicitly tested rather than assumed. The populations used to generate healthy reference data and train disease models must therefore be clearly stated when performing gut microbiome analyses, especially when intended for clinical use.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41591-018-0164-x.

Received: 23 February 2018; Accepted: 24 July 2018;
Published online: 27 August 2018

References

1. Lynch, S. V. & Pedersen, O. The human intestinal microbiome in health and disease. *N. Engl. J. Med.* **375**, 2369–2379 (2016).
2. Gilbert, J. A. et al. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94–103 (2016).
3. Gilbert, J. A. et al. Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
4. Dubinsky, M. & Braun, J. Diagnostic and prognostic microbial biomarkers in inflammatory bowel diseases. *Gastroenterology* **149**, 1265–1274.e3 (2015).
5. Konstantinov, S. R., Kuipers, E. J. & Peppelenbosch, M. P. Functional genomic analyses of the gut microbiota for CRC screening. *Nat. Rev. Gastroenterol. Hepatol.* **10**, 741–745 (2013).
6. Routy, B. et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91–97 (2018).
7. Bouter, K. E., van Raalte, D. H., Groen, A. K. & Nieuwdorp, M. Role of the gut microbiome in the pathogenesis of obesity and obesity-related metabolic dysfunction. *Gastroenterology* **152**, 1671–1678 (2017).
8. Tang, W. H. et al. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* **368**, 1575–1584 (2013).
9. Graham, C., Mullen, A. & Whelan, K. Obesity and the gastrointestinal microbiota: a review of associations and mechanisms. *Nutr. Rev.* **73**, 376–385 (2015).
10. Wieland, A., Frank, D. N., Harnke, B. & Bambha, K. Systematic review: microbial dysbiosis and nonalcoholic fatty liver disease. *Aliment. Pharmacol. Ther.* **42**, 1051–1063 (2015).
11. Forslund, K. et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
12. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
13. Martínez, I. et al. The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. *Cell Reports* **11**, 527–538 (2015).
14. Gupta, V. K., Paul, S. & Dutta, C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front. Microbiol.* **8**, 1162 (2017).
15. 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).
16. Costea, P. I. et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
17. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
18. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
19. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
20. Anderson, M. J. & Walsh, D. C. I. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecol. Monogr.* **83**, 557–574 (2013).
21. Zhou, Y. et al. Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. *mSystems* **3**, e00188–e00117 (2018).
22. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
23. Walters, W. A., Xu, Z. & Knight, R. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett.* **588**, 4223–4233 (2014).
24. McDonald, D. et al. American Gut: an open platform for citizen science microbiome research. *mSystems* **3**, e00031–18 (2018).
25. Stagaman, K. et al. Market integration predicts human gut microbiome attributes across a gradient of economic development. *mSystems* **3**, e00122–17 (2018).
26. Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. & Relman, D. A. The application of ecological theory toward an understanding of the human microbiome. *Science* **336**, 1255–1262 (2012).
27. Liao, M. et al. Comparative analyses of fecal microbiota in Chinese isolated Yao population, minority Zhuang and rural Han by 16sRNA sequencing. *Sci. Rep.* **8**, 1142 (2018).

Acknowledgements

We acknowledge the contributions of the 308 local CDC investigators and registered nurses for their help with collection point maintenance and metadata and stool sample collection. We thank all volunteers who participated in this project. This study was supported by the National Projects of Major Infectious Disease Control and Prevention (2017ZX10103011 (H.W.Z.)), National Natural Science Foundation of China (NSFC31570497 (H.W.Z.), 31322003 (H.W.Z.), and 81671171 (J.Y.)), China Postdoctoral Science Foundation (C1090132 (Y.H.)) and the Science and Technology Planning Project of Guangdong Province, China (2015A030401055 (W.W.)).

Author contributions

Y.H., W.W., W.-J.M. and H.-W.Z. designed the study; W.W., C.-Z.H., Z.-H.C, G.-Y.J., Y.-J.X. and W.-J.M. organized data collection and trained local CDC investigators; M.-X.C., Z.-D.-X.Z., P.M., X.-J.C., Z.-H.R., L.-Y.L. and N.Y. processed the samples; Y.H., H.-M.Z., P.L., H.-F.S., X.W., C.-B.W., P.C., J.Y. and H.-W.Z. analyzed the data; Y.H., D.M., W.-J.M., R.K. and H.-W.Z. drafted the manuscript; and R.K. and J.R. offered advice regarding the design of the study, data analysis and manuscript writing.

Competing interests

L.-Y.L., X.W. and C.-B.W. are employees of Shenzhen Fun-Poo Biotech Co., LTD.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0164-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to W.-J.M. or H.-W.Z.

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

Methods

Description of study design and population. This study was conducted in south China using data obtained from 14 randomly selected districts or counties in Guangdong province. The districts and counties ranged from highly developed metropolises (for example, Nanshan district, Shenzhen) to rural counties. In each district or county, three neighborhoods or townships were selected, and in each neighborhood or township, two communities or villages were selected. Selection was random and performed using probability proportional to size (PPS) sampling. In each community or village, 45 households were randomly selected, and family members aged 18 years old or older were investigated after agreeing to participate and signing the informed consent form. Family members were not included in this investigation if they (i) had lived in their household for less than 6 months of the year before the investigation; (ii) did not live in the household during the investigation; (iii) were pregnant or (iv) unable to participate because of critical illness, disability or hospitalization or (v) were not willing to participate for other reasons. In each community or village, a collection point was equipped with −20 °C refrigerators, and local Center for Disease Control (CDC) faculty were trained to conduct the investigation. In all, 14,102 individuals were approached, 9,172 individuals were investigated, and 7,009 participants were included in the final analysis after all individuals who did not properly donate their stool or blood samples or complete their questionnaire or whose stool sample sequences provided < 10,000 reads were excluded. The age range of the studied population was 18–97 years old, and the mean age ± s.d. was 52.78 ± 14.68 years.

Stool sample collection and transportation. Participants were given a stool sampler, ice bag and ice box and were provided detailed illustrated instructions describing how to collect and store the samples. The instructions stated that the ice bag should be stored in the participant's refrigerator (−18 to −20 °C) at home until sample collection. After defecation, each participant collected their stool sample, recorded its Bristol stool score in the stool sampler and stored the sample with the prestored ice bag in the ice box as instructed. Participants were required to send the stool sample to a nearby collection point as soon as possible within 1 d. For samples taken at night, it was recommended that the stool sample should be sent to the collection point the next morning. Stool samples that were not delivered to the collection point within 24 h were discarded. Most of the stool samples received at all sampling sites were received in the morning. Collected samples were stored in a freezer (−18 to −20 °C) at the collection point for less than 3 d and then were transported to the research laboratory (Guangdong CDC) by Kingmed Company (Guangzhou, China) in a cold-chain vehicle in which the temperature was monitored during transportation. The village and community collection points were between 10 and 480 km from the research laboratory. The transportation company made 84 deliveries to the research laboratory, and transportation times ranged from 30 min to 28 h. Samples were transported and stored after arrival at the research laboratory in −80 °C freezers until they were further processed.

Wet lab processing of stool samples. Before the specimens were submitted to laboratory analysis, we prepared external standards to control for potential batch effects because multiple technicians and machines were involved in sample processing. Briefly, fecal samples were collected from three donors. For each donor, the samples were manually homogenized to obtain an even mixture, divided into 200 tubes and stored at −80 °C.

Total bacterial DNA was extracted using a Fecal DNA Bead Isolation kit (Bioeasy, Shenzhen) according to the manufacturer's instructions on a KingFisher FLEX instrument for automatic processing. Identical protocols were followed for all samples, including the addition of three external standards to each batch. The 16 S rRNA gene V4 region was amplified by the following barcoded primers (shown from 5′ to 3′): V4F, GTGYCAGCMGCCGCGGTAA and V4R, GGACTACNVGGGTWTCTAAT[28]. The barcodes were designed using Barcrawl[29] so that there were 3-bp differences between any two barcodes. The following three PCR machines were used in the present study: ABI-1:ViiA 7 Dx, ABI-2:ViiA 7 and Bioneer. They were operated by two professional technicians. The following PCR conditions were used: initial denaturation at 94 °C for 5 min followed by 30 cycles of denaturation at 94 °C for 30 s, annealing at 52 °C for 30 s, extension at 72 °C for 45 seconds and a final extension at 72 °C for 5 min. All PCR products were sequenced on an Illumina Hiseq 2500 using 500 cycle version 2 reagent kits (Beijing Genome Institute, BGI; Beijing, China).

To compare the results obtained from the DNA extraction kit used in the present study to those obtained using a Mobio PowerSoil DNA extraction kit (used in the American Gut Project[30]), we selected 18 samples from our study and processed them simultaneously using a Fecal DNA Bead Isolation kit and a Mobio PowerSoil DNA extraction kit. The results were then compared.

Next-generation sequencing protocol. *The method of sample detecting.* Sample testing included concentration, sample integrity and purification. Concentration was detected by fluorometer or microplate reader (for example, QubitFluorometer, Invitrogen). Sample integrity and purification was detected by Agarose Gel Electrophoresis (concentration of agarose gel: 1%; voltage: 150 V; electrophoresis time: 40 min).

Library construction. A 10-μg DNA sample was fragmented by Covaris, and the fragmented DNA was detected through gel electrophoresis.

End repair. After purification of fragmented DNA by QIAquick PCR purification kit (Qiagen), the fragmented DNA and end-repair mix was combined, and the end-repair reaction was incubated for 30 min at 20 °C. Then the end-repaired DNA was purified with the QIAquick PCR purification kit.

A-tailing. The end-repaired DNA and A-tailing mix were combined, and the A-Tailing reaction was incubated for 30 min at 37 °C. Then the end-repaired DNA was purified with the QIAquick PCR Purification Kit.

Adaptor ligation. The adenylate 3′ ends DNA, PCR-free index adapters and T4 DNA ligase mix were combined and mixed well by pipetting. The ligate reaction was incubated for 15 min at 20 °C. Purify the end-repaired DNA with the QIAquick PCR purification kit.

Size-selected library and products purification. The ligation products are selected by running a 2% agarose gel to recover the target fragments. Purify the gel with QIAquick gel extraction kit (QIAGEN).

Validation of the library. The final library was quantitated in two ways: the average molecule length was determined using the Agilent 2100 bioanalyzer instrument (Agilent DNA 1000 Reagents), and the library was quantified by real-time qPCR (TaqMan Probe).

Sequencing libraries. The qualified libraries were amplified on cBot to generate the cluster on the flowcell (HiSeq PE Cluster Kit v4 cBot, Illumina). The amplified flowcell was sequenced pair-end on the HiSeq2500 System (HiSeq SBS Kit V4, Illumina).

Metadata collection. The collected metadata covered participants' sociodemographic features, anthropometric features, and information related to lifestyle, diet, medication, diseases, and biochemistry. In all, 72 factors were selected and further categorized into the following seven groups: 7 demographic factors, 23 diseases occurring in at least 18 individuals, 7 physiological factors, 7 blood parameters, 22 lifestyle factors, including 19 related to dietary information and 3 related to medication, and Bristol stool scores. A description of each factor is listed in Supplementary Table 1.

Participants were asked to fast for 12 h before blood collection. Fasting blood samples were collected by registered nurses, anticoagulated with EDTA-K2, and stored and transported in the same method used for the stool samples. FBG, TG, TCHO, high-density lipoprotein (HDL), low-density lipoprotein (LDL), UA and hemoglobin (Hb) levels were measured on a Hitachi 7600 automatic biochemical analyzer using reagents obtained from Wako Pure Chemical Industries Ltd. at the National CDC of China. Anthropometric data included height, weight, waist, SBP and DBP. Additionally, heart rate was measured under fasting conditions (at the same time as blood sample collection) by three trained faculty members, and average values were calculated. Blood pressure was measured using an OMRON HBP1300 electronic sphygmomanometer at 1-min intervals.

In face-to-face questionnaire interviews, we collected host metadata, including sociodemographic features and data related to lifestyle, diet, medication and diseases. Self-reported diseases were confirmed by analyzing medical histories from district-level or higher hospitals. All information provided during the interviews was recorded, and the interviews were performed by two CDC faculty members.

MetS was defined according to the Joint Committee for Developing Chinese Guidelines on the Prevention and Treatment of Dyslipidemia in Adults as meeting three of the five following criteria: (i) waist > 90 cm (male) or waist > 85 cm (female), (ii) FBG ≥ 6.1 mmol/L (110 mg/dl) or previously diagnosed with diabetes mellitus, (iii) TG ≥ 1.7 mmol/L (150 mg/dl), (iv) HDL < 1.04 mmol/L (40 mg/dl) and (v) SBP/DBP ≥ 130/85 mm Hg or previously diagnosed with high blood pressure. Type 2 diabetes mellitus (T2DM) was determined by self-report (conformed with medical history) or FBG ≥ 7.0 mmol/L. Obesity was defined as a BMI ≥ 28 kg/cm^2.

Bioinformatics and biostatistics. Approximately 600 million paired-end raw sequences were obtained from BGI, which had been partitioned by the sequencing center into 18 pair-end fastq files for convenient data transfer and processing. For each pair-end fastq, we first trimmed each sequence to 200 bp from the 5′ end and used SeqPrep (implemented in Quantitative Insights Into Microbial Ecology, QIIME[18]) to merge the pair-end sequences based on approximately 100 bp of overlap between the two pair-end sequences. The merged fastq file (spanning the entire 16 S rRNA gene V4 region) was quality controlled by QIIME (1.9.1), and the quality scores were then cast away. A laboratory-developed script was used to split the fasta files according to the pair-end barcode information; this was based on the principle that the barcodes at both ends should exactly match one sample and that no errors should be allowed in the barcode and primer region (the bioinformatics procedure used to process raw sequences to an OTU table is documented in

https://github.com/SMUJYYXB/GGMP-Regional-variations). We used a QIIME workflow script, pick_closed_reference_otus.py, to perform referenced-based OTU clustering and used USEARCH6.1 as the reference mode against the GreenGenes database v13_8 (refs. [31–33]) and for BIOM generation (BIOM, the biological observation matrix format[34]). A default similarity threshold was used for clustering, which was defined as 97%. Phylogenetic relationships and the taxonomy of representative OTU sequences were determined using the GreenGenes database in default. Finally, we merged all BIOM files for each split fastq for downstream biostatistical analysis.

To calculate the variation explained by each of our collected host factors, we performed an Adonis test implemented in QIIME. Each host factor was calculated according to its explanation rate, and *P* values were generated based on 1,000 permutations. All *P* values were then adjusted using the Benjamini–Hochberg method. To validate the robustness of the geographical effect, an additional four statistical methods were performed. Anosim and MRPP were performed in QIIME using default parameters. The db-RDA was performed in R (version 3.2.2) using the 'capscale' function in the 'vegan' R package. The 'envfit' function in the 'vegan' R package was also used. This function performs MANOVA and linear correlations for categorical and continuous variables, respectively.

To build microbiome models to identify the origin of fecal samples, a random forest classification model was trained to predict districts using the 7,009 fecal samples obtained for this project. This analysis was performed in the R language using the 'caret' (version 6.0.68) and 'randomForest' (version 4.6.12) packets. The parameters were set as follows: ntree = 5,000 and the default mtry of sqrt(p), where p is the number of input 97%-identity OTUs (features). The AUCs for each district were calculated by receiver operating characteristic (ROC) analysis using a tenfold cross-validation technique in the training set.

To identify microbiome signatures for T2DM, MetS, obesity and fatty liver, we performed multivariate association analyses (MaAsLin) to identify associations between OTUs and each of the four diseases[19,35]. We performed local multivariate association with linear models (MaAsLin version 0.0.2, downloaded from https://bitbucket.org/biobakery/maaslin/) in R (version 3.2.2). MaAsLin is a multivariate statistical framework that identifies associations between continuous and discrete clinical metadata and microbial community abundances. Briefly, the percentage of each OTU was arcsin-square-root-transformed. For each correlation between metadata (as predictors) and transformed OTU abundances (as response variables), we used confounders (such as age, gender, Bristol stool scale and geographic location[17,19,36]) as covariates. To further reduce the computational load, OTUs that were present in <10% of the population were excluded, leaving 950 OTUs that constituted 96% of the initial abundance in the following analyses. The Benjamini– Hochberg method was used to adjust *P* values for multiple hypotheses (q values).

To determine whether disease models built in one district based on gut microbiota can be cross-applied to other districts, we built machine learning models of metabolic diseases, including T2DM, MetS, obesity and fatty liver, based on the gut microbiota profiles obtained in each district to evaluate their cross-applicability. In each district and for each constructed disease model, candidate OTUs for modeling were first selected as follows: a Wilcoxon rank sum test was performed to compare healthy individuals with patients, and OTUs with $P < 0.1$ were retained for machine-learning construction. Then, a machine-learning model was built for a specific disease in the same district using the same R packages and parameters described above. The interpolated AUCs were evaluated using a tenfold cross-validation technique in the training set and then applied to the other districts to generate AUCs for cross-applications. Paired nonparametric *t*-tests were used to determine the significance of differences between the AUCs of self-validation and cross-application.

To illustrate our proposed principle that it is easier to interpolate within a smaller scale than to extrapolate (with interpolation within a larger scale lying between), we calculated interpolation and extrapolation AUCs on different geographical levels in the MetS model. Extrapolations between communities were performed within the same neighborhood or between neighborhoods within the same district depending on the level of analysis. The modeling procedure was performed in accordance with the above descriptions.

To determine the extrapolation efficiency of our province-level model of MetS in other populations, we reprocessed a dataset from a Korean population study of MetS[37] using the same bioinformatics pipeline as for our own study. The modeling and extrapolating procedure were in performed per the descriptions above.

Analysis of icebox transportation on the level of Proteobacteria in fecal samples. Proteobacteria can overgrow if fecal samples are not stored or transported at a low temperature. Because our samples were initially stored in participants' homes and transferred to an icebox within 24 h, we examined whether Proteobacteria blooming may have occurred during this period. First, we determined the level of Proteobacteria in 20 local individuals' samples that were not stored and transported in an icebox but were instead directly processed. We collected fecal samples from an additional 20 individuals with no self-reported diseases at the Physical Examination Center of Zhujiang Hospital, Guangzhou, Guangdong, China (PEC). The average age of these 20 individuals was 55.35 years old, which was not significantly different from that of the population in Guangzhou studied in this project. We transported their samples to a refrigerator kept at −80 °C within 1 h after collection and compared their Proteobacteria level with that in GGMP. The results showed that the level of Proteobacteria was comparable to that observed in the GGMP ($P = 0.879$, also see Supplementary Fig. 2a). Second, we compared fecal microbiota profiles between parallel samples that were either immediately processed or processed after being stored in an icebox for 24 h. We collected fecal samples from two individuals. The samples obtained from each individual were divided into 12 parallel samples, 6 of which were processed immediately; the other 6 samples were stored in three iceboxes for 24 h and then processed. The results showed that the temperature inside the icebox was approximately 4 °C after 24 h (Supplementary Table 14) and that the relative abundance of Proteobacteria was not significantly different between the samples that were processed at different storage times ($P = 0.44$; Supplementary Fig. 2b). After temporary storage and transportation in an icebox, our samples were stored in refrigerators and transported using cold-chain vehicles, and we therefore assume no blooming of Proteobacteria occurred during this process.

Analysis of external standard samples. In each batch of extracted DNA, we added one external standard sample (ESS) from each of the three donors. All ESSs were processed in the same way as our experimental samples in the bioinformatics analysis. We performed principle coordinate analysis (PCoA) to validate that the ESSs were clustered according to their donors or according to other batch effects, for example, PCR operators or PCR machines, because two technicians performed the PCRs on three PCR machines. We found that the ESSs were clustered according to their donors rather than their PCR operators or PCR machines (Supplementary Fig. 2c). This finding indicated that the experimentally induced variations were much smaller than the interindividual variations. Therefore, comparisons between individuals were unlikely to be affected by experimental batch effects.

Comparisons between DNA extraction kits. To determine whether the higher amount of Proteobacteria we observed was due to technical bias introduced by the DNA extraction kit, we selected 18 samples and processed them with a Bioeasy Fecal DNA bead Isolation kit (in the present study referred to as Bioeasy) and a Mobio Powersoil kit (a popular kit used in other projects, including the American Gut Project and the Earth Microbiome Project, and referred to here as Mobio). We first compared profiles from the two kits using PCoA to determine whether the samples could be separated on the basis of unsupervised classifications. The gut microbiomes profiled by the two different kits could clearly be separated based on unweighted UniFrac distances, which reflect only the presence or absence of information related to OTUs and the phylogenetic relationships among OTUs. However, distances evaluated by considering the abundances of OTUs, for example, weighted UniFrac or Bray Curtis, did not separate the samples according to which processing kit was used (Supplementary Fig. 2d). This finding indicated that the gut microbiomes produced by the two different DNA extraction kits might differ in the presence and/or absence of rare OTUs but not in the abundances of more prevalent OTUs. We then compared the bacterial compositions captured by the two kits (Supplementary Fig. 2e). We found that Bioeasy tended to capture fewer Bacteroidetes (including significantly fewer *Bacteroides*) and more Firmicutes (including significantly more *Faecalibacterium*) species. Moreover, the abundances of Proteobacteria were similar between the two kits (8.67% for Bioeasy versus 8.03% for Mobio; Wilcoxon rank-sum test, $P > 0.05$). This finding indicated that the relative abundances of certain taxonomies generated by the two kits can differ, but Proteobacteria was much less affected than were Bacteroidetes or Firmicutes. We then used the Procrustes algorithm to determine whether the relationships between the samples captured by the two kits were similar (Supplementary Fig. 2f). Briefly, Procrustes compares two PCoAs (the same samples generated using different methods) by rotating and scaling one of them and seeing how similar the patterns are. The relationships captured by the two kits were highly and significantly similar to each other. This finding indicated that the results of comparisons between the samples were highly identical regardless of which kit was used even when they differed in their captured microbiome profiles.

Ethics approval and consent to participate. The present study was approved by the Ethical Review Committee of the Chinese Center for Disease Control and Prevention under Approval Notice No. 201519-A. Written consent was obtained from all participants.

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

Code availability. The bioinformatic code used for raw data processing is available from https://github.com/SMUJYYXB/GGMP-Regional-variations.

Data availability. The raw data for 16 S rRNA gene sequences are available from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/) at accession number PRJEB18535. The metadata are available in Supplementary Table 13.

References

28. Walters, W. et al. Improved bacterial 16S rRNA Gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *mSystems* **1**, e00009–e00015 (2015).
29. Frank, D. N. BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinformatics* **10**, 362 (2009).
30. McDonald, D., Birmingham, A. & Knight, R. Context and the human microbiome. *Microbiome* **3**, 52 (2015).
31. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
32. McDonald, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME. J.* **6**, 610–618 (2012).
33. He, Y. et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* **3**, 20 (2015).
34. McDonald, D. et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* **1**, 7 (2012).
35. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome. Biol.* **13**, R79 (2012).
36. Xie, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584 (2016).
37. Lim, M. Y. et al. The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. *Gut* **66**, 1031–1038 (2017).

nature research

Corresponding author(s): Hong-Wei Zhou

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

No statistical method was applied for sample size determination because such a method is not available for gut microbiota observational study to date. Sample sizes for microbiota studies could vary from tens to thousands. As one of the largest gut microbiota observational study, we have involved 7,009 participants in the final analysis, with an average of 500 in each of the 14 districts we have surveyed, to insure that normal variations in gut microbiota can be observed.

2. Data exclusions

Describe any data exclusions.

In each community/village, 45 households were randomly selected, and family members aged 18 years old or older were investigated after agreeing to participate and signing the informed consent form. Family members were not included in this investigation if they (1) had lived in their household for less than six months of the year before the investigation; (2) did not live in the household during the investigation; (3) were pregnant or (4) unable to participate due to critical illness, disability or hospitalization or (5) not willing to participate for other reasons. In each community/village, a collection point was equipped with -20°C refrigerators, and local CDC (Center for Disease Control) faculty were trained to conduct the investigation. In all, 14102 individuals were approached, 9172 individuals were investigated, and 7009 participants were included in the final analysis after all individuals who did not properly donate their stool or blood samples or complete their questionnaire or whose stool sample sequences provided less than 10,000 reads were excluded.

When calculating the associations between gut operational taxonomic unites (OTUs) and metabolic diseases, we excluded any patients who were taking medication to control blood glucose, because it was reported that metformin significantly impact gut microbiota. To further reduce the computational load, OTUs that were present in less than 10% of the population were also excluded.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

Experimental replication was not attempted

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

14 districts/counties were chosen at random from Guangdong province China. In each district/county, three neighbourhoods/townships were selected, and in each neighbourhood/township, two communities/villages were selected randomly using PPS (Probability Proportional to Size) sampling. In each community/village, 45 households were randomly selected and approached for investigation.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

The presented observational study characterizes normal variations in human gut microbiota. Blinding is not relevant.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Data analysis was performed using QIIME (1.9.1) and R (3.2.2). Machine learning modeling was conducted using R packages caret (version 6.0.68) and randomForest (version 4.6.12). Multivariate analysis with linear model was performed using MaAsLin (0.0.2). Code were deposited in GitHub: https://github.com/SMUJYYXB/GGMP-Regional-variations.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	No unique materials were used in the present study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

We conducted a cross-sectional study, collected fecal samples from the community population and did not aim at any specific diseases. All 7,009 volunteers in the present analysis were from Guangdong, China and distributed in 14 districts. The age range of the studied population is 18-97 years old, and the mean±SD of age is 52.78±14.68. The population included 3161 males and 3848 females. In our population, metabolic diseases were the most prevalent ones among other diseases, and our population included 608 type two diabetes mellitus, 163 fatty liver, 1404 metabolic syndrome and 707 obese individuals.