A transomic cohort as a reference point for promoting a healthy human gut microbiome

Jie, Zhuye; Liang, Suisha; Ding, Qiuxia; Li, Fei; Tang, Shanmei; Wang, Dan; Lin, Yuxiang; Chen, Peishan; Cai, Kaiye; Qiu, Xuemei; Li, Qiang; Liao, Yunli; Zhou, Dongsheng; Lian, Heng; Zuo, Yong; Chen, Xiaomin; Rao, Weiqiao; Ren, Yan; Wang, Yuan; Zi, Jin; Wang, Rong; Zhou, Hongcheng; Lu, Haorong; Wang, Xiaohan; Zhang, Wei; Zhang, Tao; Xiao, Liang; Zong, Yang; Liu, Weibin; Yang, Huanming; Wang, Jian; Hou, Yong; Liu, Xiao; Kristiansen, Karsten; Zhong, Huanzi; Jia, Huijue; Xu, Xun

Published in:
Medicine in Microecology

DOI:
10.1016/j.medmic.2021.100039

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
A transomic cohort as a reference point for promoting a healthy human gut microbiome

Zhuye Jia, Suisha Liang, Qiuxia Ding, Fei Li, Shanmei Tang, Dan Wang, Yuxiang Lin, Peishan Chen, Kaiye Cai, Xuemei Qiu, Qiang Li, Yunli Liao, Dongsheng Zhou, Heng Lian, Yong Zuo, Xiaomin Chen, Weiqiao Rao, Yan Ren, Yuan Wang, Jin Zhi, Rong Wang, Hongcheng Zhou, Haorong Lu, Xiaohan Wang, Wei Zhang, Tao Zhang, Liang Xiao, Yang Zong, Weibin Liu, Huanming Yang, Jian Wang, Yong Hou, Xiao Liu, Karsten Kristiansen, Huanzi Zhong, Huijue Jia, Xun Xu

A B S T R A C T

More than a decade of gut microbiome studies have a common goal of improving human health. However, while most of the disease studies have focused on the elderly or the middle-aged, a reference cohort for the gut microbiome in young individuals has been lacking. It is also not clear what other omics data need to be measured to better understand the gut microbiome. Here, we present a cohort including 2183 adults with high-depth metagenomic shotgun sequencing data for the fecal microbiome and other omics data. In this multiomic cohort, we observe a number of vitamins, hormones, amino acids, and trace elements that correlated with the gut microbiome. Many of the associations are validated in an additional cohort consisting of 1404 individuals. Our comprehensive data are poised to provide advice to future populations and mechanistic study designs to better understand and manage our gut microbiome.

1. Introduction

The gut microbiome has been implicated in a growing list of complex diseases, showing great potential for the diagnosis and treatment of metabolic, autoimmune, and neurological diseases as well as cancer. While case-control studies have been illuminating [1], recently published studies have emphasized the difficulty in extrapolating to natural cohorts due to heterogeneity in location and ethnicity [2,3]. Thus far, only a few cohorts made use of metagenomic shotgun sequencing instead of 16S rRNA gene amplicon sequencing, the largest being the LifeLines Deep cohort (n = 1135 and 32 million reads per sample) from the Netherlands [4–7]. Fecal or plasma metabolites are more or less included in gut microbiome studies, but the conclusions usually did not go beyond short-chain fatty acids (SCFA), amino acids, the vitamin B complex, or bile acids. Levels of trace elements such as arsenic have been a health concern [8,9] but have not been studied in a human cohort with microbiome data. Exercise [10] and lifestyle have been reported to modulate the microbiota, but all these factors have not been studied and compared in the same cohort with the same baseline. In addition, biological sex is a strong determinant for the gut microbiome in mice and livestock [11–13]. The potential impact of hormones on the human gut microbiome or vice versa, remains unclear.

As part of the 4D-SZ (transomic, with more time points in future studies, from Shenzhen, China) cohort, here we present metagenomic
shotgun sequencing data sufficient for high-resolution taxonomic and functional profiling (86.15 ± 23.35 million reads per sample) of the fecal microbiome in a cohort of 2183 adults, along with questionnaire data, physical fitness tests, and plasma metabolome. Transomics data in this Han Chinese cohort were analyzed in the context of the fecal microbiome and uncovered previously overlooked measurements such as aldosterone, testosterone, trace elements, and vitamin A that influence the gut microbiome. The results were validated in an additional cohort with 1400 individuals. Thus, the 4D-SZ cohort could provide clues for the inclusion of overlooked omics data, including measurements of trace element from the environment, providing more comprehensive yet streamlined designs of questionnaires to better understand how each individual's gut microbiome came about and was influenced by the environment (Supplementary Figs. 1 and 2).

2. Materials and methods

2.1. Study cohort

As a part of 4D-SZ, all the >2000 volunteers for the first cohort were recruited between May 2017 and July 2017 during a physical examination in Shenzhen. The 1400 volunteers for the second cohort were also recruited in 2017 until late September in multiple cities in China, including Wuhan, Qingdao, and Tianjin. The samples for each omics are shown in Supplementary Table 1c. Baseline characteristics of the cohort are shown in Supplementary Table 1b and d.

The study was approved by the Institutional Review Boards (IRB) at BGI-Shenzhen (BGI-IRB19121), and all participants provided written informed consent at enrolment.

2.2. Demographic data collection

The lifestyle questionnaire contained 56 entries involving age, marital status, disease history of the volunteer and his/her family, and eating and exercise habits (Supplementary Table 1b and d). The psychological questionnaire contained 18 entries to evaluate irritability, dizziness, frustration, fear, appetite, self-confidence, and resilience (Supplementary Table 1b).

2.3. Fecal sample collection

Fecal samples were self-collected by the volunteers, using a kit containing a room temperature stabilizing reagent to preserve the metagenome [14], which can preserve the fecal metagenome for at least two weeks at room temperature. The samples were frozen at −80 °C on the same day and were only thawed for DNA extraction. The overnight fasting blood samples were drawn from a cubital vein of volunteers by the medical doctors.

2.4. DNA extraction and metagenomics shotgun sequencing

DNA extraction of the stored fecal samples within few months after collection was performed as previously described [15]. This manual extraction protocol has been described in detail before [16] and used in many other studies [17–20]. Briefly, a frozen aliquot of fecal sample was suspended in 250 μl of guanidine thiocyanate, 0.1 M Tris (pH 7.5), and 40 μl of 10 % N-lauroyl sarcosine. Then, 500 μl (5 %) of N-lauroyl sarcosine were added. After 1 h of incubation, 500 μl of glass beads (0.1 mm) and 500 μl of TENP were added in the tube for vortexing and centrifuged. The supernatant was transferred to a new tube, and DNA was precipitated by isopropanol [14]. Metagenomic sequencing performed on the BGISEQ-500 platform (PCR-free without size selection, 100 bp of single-end sequencing) and quality-controlled were conducted as previously reported [21].

2.5. Amino acid measurements

Plasma of 40 μl was deproteinized with 20 μl (10 %) (w/v) of sulfosalicylic acid (Sigma) containing internal standards, and 120 μl aqueous solution was then added. After centrifugation, the supernatant was used for analysis. The analysis was performed using ultra-high-pressure liquid chromatography (UHPLC) coupled with an AB Sciex Qtrap 5500 mass spectrometer (AB Sciex, US) with the electrospray ionization (ESI) source in positive ion mode. A Waters ACQUITY UPLC HSS T3 column (1.8 μm, 2.1 × 100 mm) was used for amino compound separation with a flow rate at 0.5 ml/min and column temperature of 55 °C. The mobile phases were (A) water containing 0.05 % and 0.1 % formic acid (v/v) and (B) acetonitrile containing 0.05 % and 0.1 % formic acid (v/v). The gradient elution was 2 % B kept for 0.5 min, then changed linearly to 10 % B during 1 min, continued up to 35 % B in 2 min, increased to 95 % B in 0.1 min, and maintained for 1.4 min. Multiple reaction monitoring (MRM) was used to monitor all amino compounds. The mass parameters were as follows: curtain gas flow 35 L/min, collision gas (CAD) was medium, ion source gas 1 (GS 1) flow 60 L/min, ion source gas 2 (GS 2) flow 60 L/min, ionspray voltage (IS) 5500 V, and temperature 600 °C. All amino compound standards were purchased from Sigma and Toronto Research Chemical (TRC).

2.6. Hormone measurements

Plasma of 250 μl was diluted with a 205 μl aqueous solution. For SPE experiments, HLB (Waters, USA) was activated with 1.0 ml of dichloromethane, acetonitrile, and methanol, respectively and was equilibrated with 1.0 ml of water. The pretreated plasma sample was loaded onto the cartridge and was extracted using gravity. Clean up was accomplished by washing the cartridges with 1.0 ml of 25 % methanol in water. After drying under vacuum, samples on the cartridges were eluted with 1.0 ml of dichloromethane. The eluted extract was dried under nitrogen, and the residue was dissolved with 25 % methanol in water and transferred to an autosampler vial prior to LC–MS/MS analysis. The analysis was performed using UHPLC coupled with an AB Sciex Qtrap 5500 mass spectrometer (AB Sciex, US) with the atmospheric pressure chemical ionization (APCI) source in positive ion mode. A Phenomenex Kinetex C18 column (2.6 μm, 2.1 × 50 mm) was used for steroid hormone separation with a flow rate at 0.8 ml/min and column temperature of 55 °C. The mobile phases were (A) water containing 1 mM ammonium acetate and (B) methanol containing 1 mM ammonium acetate. The gradient elution was 25 % B kept for 0.9 min, then changed linearly to 40 % B during 0.9 min, continued up to 70 % B in 2 min, increased to 95 % B in 0.1 min, and maintained for 1.6 min. MRM was used to monitor all steroid hormone compounds. The mass parameters were as follows: curtain gas flow 35 L/min, collision gas (CAD) was medium, ion source gas 1 (GS 1) flow 60 L/min, ion source gas 2 (GS 2) flow 60 L/min, ionspray voltage (IS) 5500 V, and temperature 600 °C. All steroid hormone compound standards were purchased from Sigma, TCR, Cerilliant, and Dr. Ehrenstorfer.

2.7. Trace element measurements

Whole blood (200 μl) was transferred into a 15 ml polyethylene tube and diluted 1:25 with a diluent solution consisting of 0.1 % (v/v) Triton X-100, 0.1 % (v/v) HNO3, 2 mg/l AU, and internal standards (20 μg/l). The mixture was sonicated for 10 min before ICP-MS analysis. Multi-element determination was performed on an Agilent 7700x ICP-MS (Agilent Technologies, Tokyo, Japan) equipped with an octopole reaction system collision/reaction cell technology to minimize spectral interferences. The continuous sample introduction system consisted of an autosampler, a quartz torch with a 2.5-mm diameter injector with a Shield Torch system, and a Scott double-pass spray chamber and nickel cones (Agilent Technologies, Tokyo, Japan). A glass concentric MicroMist nebulizer (Agilent Technologies, Tokyo, Japan) was used to analyze diluted samples.
2.8. Water-soluble vitamins measurements

Plasma (200 μl) was deproteinated with 600 μl methanol (Merck), water, acetic acid (9:1:9:1) containing internal standards, thiamine-(4-methyl-13C-thiazol-5-yl-13C3) hydrochloride (Sigma-Aldrich), levomefolic acid-13C3, d3, riboflavin-13C, 15N2, 4- pyridoxic acid-d3 and pantothenic acid-13C3, and 15 N hemi calcium salt (Toronto Research Chemicals). The supernatant (500 μl) was dried by nitrogen flow. Water (60 μl) was added to the residues, vortexed, the mixture was centrifuged, and the supernatant was used for analysis. The analysis was performed by UPLC coupled to a Waters Xevo TQ-S Triple Quad mass spectrometry (Waters, USA) with the ESI source in positive ion mode. A Waters ACQUITY UPLC HSS T3 column (1.7 μm, 2.1 × 50 mm) was used for water-soluble vitamins separation with a flow rate at 0.45 ml/min and column temperature of 45 °C. The mobile phases were (A) 0.1% formic acid in water and (B) 0.1 % formic acid in methanol. The following elution gradient was used: 0–1 min, 99.0%-99.0 % A; 1–1.5 min, 99.0 % A–97.0 % A; 1.5–2 min, 97.0 % A–70.0 % A; 2–3.5 min, 70 % A–30 % A; 3.5–4.0 min, 30 % A–10.0 % A; 4.0–4.8 min, 10 % A–1.0 % A; 4.9–6.0 min, and 99.0 % A–99.0 % A. Multiple reaction monitoring (MRM) was used to monitor all water-soluble vitamins. The mass parameters were as follows: the capillary voltages of 3000 V and source temperature of 150 °C were adopted. The desolvation temperature was 500 °C. The collision gas flow was set at 0.10 ml/min. The cone gas and desolvation gas flow were 150 l/h and 1000 l/h, respectively. All water-soluble vitamin standards were purchased from Sigma-Aldrich (USA).

2.9. Fat-soluble vitamin measurements

Plasma (250 μl) was deproteinized with 1000 μl methanol and acetonitrile (v/v, 1:1, Fisher Chemical) containing internal standards, all-trans-Retinol-d5, 25-HydroxyVitamin-D2-d6, 25-HydroxyVitamin-D3-d6, vitamin K1-d7, and α-Tocopherol-d6 (Toronto Research Chemicals). The supernatant (900 μl) was dried by nitrogen flow. Acetonitrile (80 μl, 80 %) was added to the residues, vortexed, the mixture was centrifuged, and the supernatant was used for analysis. The analysis was performed by UPLC coupled with an AB Sciex Qtrap 4500 mass spectrometry (AB Sciex, USA) with the APC/PCI source in positive ion mode. A Waters ACQUITY UPLC BEH C18 column (1.7 μm, 2.1 × 50 mm) was used for fat-soluble vitamins separation with a flow rate at 0.50 ml/min and column temperature of 45 °C. The mobile phases were (A) 0.1 % formic acid in water and (B) 0.1 % formic acid in acetonitrile. The following elution gradient was used: 0–0.5 min, 60.0%-60.0 % A; 0.5–1.5 min, 60.0 % A–20.0 % A; 1.5–2.5 min, 20.0 % A–0.0 % A, 2.5–4.5 min, 0 % A–0.0 % A; 4.5–4.6 min, 0 % A–60.0 % A; 4.6–5.0 min, and 60.0 % A–60.0 % A. Multiple reaction monitoring (MRM) was used to monitor all fat-soluble vitamins. The mass parameters were as follows, curtain gas flow 30 l/min, collision gas (CAD) was medium, ion source gas 1 (GS 1) flow 40 l/min, ion source gas 2 (GS 2) flow 50 l/min, nebulizer current (NC) 5, and temperature 400 °C. All fat-soluble vitamin standards were purchased from Sigma-Aldrich (USA), TRC.

2.10. Medical parameters

All the volunteers were recruited during the physical examination. The medical test included blood tests, urine analysis, and routine examination of cervical secretion. The medical parameters were measured by the physical examination center and shown in Supplementary Table 1b and d InBody (InBody Co. Ltd. U.S.A.) was used to estimate body composition.

2.11. Physical fitness test

Eight kinds of physical fitness tests were performed to evaluate the physical fitness condition of the volunteers (Supplementary Table 1b). Vital capacity was measured by HK6800-FH (Hengkangjiaye, China). Eye-closed and single-legged standing was measured by HK6800-ZL. Choice reaction time was measured by HK6800-FY. Grip strength was measured by HK6800-WL. Sit and reach was measured by HK6800-TQ. Sit-ups was measured by HK6800-YW. Step index was measured by HK6800-TJ. Vertical jump was measured by HK6800-ZT. We obtained the measured result from each test. Each measured result was then scored from 1 to 5 based on its corresponding age-matched national standards (Supplementary Table 5). The direct measured result and the scores were used for analyses (Supplementary Table 2, Supplementary Table 3).

2.12. Quality control, taxonomic annotation, and abundance calculation

The sequencing reads were quality-controlled as described previously [21], and implementation of the pipeline is available at https://github.com/jiezhuye/cOLMg. Briefly, adapter trimming/filtering is automatically processed by the BGISEQ-500 sequencing platform. The raw sequences with low quality were filtered and trimmed by overall accuracy (OA) control strategy [21] using OAs1 (-Qys = 33, -minLen = 30, -Scut = 0.9, and -Qcut = 0.8). Then, the high-quality reads were aligned to hg19 by SOAP2.22 (identity ≥ 0.9) to remove human-related reads by removeHost (-D 4 -s 30 -r 1 -v 7 -i 0.9). The retained high-quality clean reads were aligned to the integrated gene catalog with 9,879,896 genes by SOAP2.22 (identity ≥ 0.95) to yield the taxonomy assignment [16]. Taxonomy of the fecal MGs/MLGs was then determined from their constituent genes as previously described [1,15,22]. The profile of an MGs/MLG was the trimmed mean (removal of the highest and lowest 5 % abundance genes) of genes abundance within this MGs/MLG. We also computed the Metaphlan2 (v2.6.0) [23] profile with the default parameter.

2.13. The factors in each type of omics predicted by another type of omics

Every factor in each omics was regressed on the relative abundances of MGs profile (found in at least 10 % of the samples) in the fecal samples using default parameters in the RFcv function from randomForest (RF) package in R (cv.fold = 5 and ntree = 500). Dichotomous variables (such as gender) and unordered categorical variables (such as region) were recoding into dummy variables. Frequency items such as yogurt eating habits were assigned to integers. RFcv R was defined as the Spearman correlation between the measured value and 5-fold cross-validation predicted value. Then, the top 5 important predictive factors in each omics type were recorded. For the prediction between gut metagenomic and other type omics, RFcv is trained on the initial cohort then applied to the validation cohort (Fig. 3). The same prediction process was done between any two types of omics. The ggplot2 package in R was then used to boxplot the predictive power of target omics factors by all kinds of other predictor omics (Fig. 2b). RFcv R (75 % quantile) between any two types of omics (from a to b and from b to a) was used to construct the bi-direction global omics correlation network using Cytoscape (Fig. 2a). R pheatmap (pheatmap) and barplot (graphics) were used to make the heatmap plot for representative factors (Fig. 2c, Supplementary Fig. 2). Mantel test (24) was done with mantel. resf function (R package ade4) for each pair of Bray-Curtis dissimilarity matrices (using vegdist function in R package vegan). To compute the p-value, 4999-times permutation was used, and 4999-times bootstrap was used to compute the 95 % confidence interval.

2.14. Statistical adjustment for factors that may influence the gut microbiome

Associations between gut microbiome MGs, functional modules, Shannon diversity, and other types of omics data were calculated by linear model or partial Spearman correlation, adjusted for potential confounders, including gender, age, BMI, health products (amino acid, vitamin, and calcium), antivirus, antibiotics, drugs (currently using antihypertensive drugs, hyperglycemic drugs, and hypolipidemic drugs),
days since last menstrual bleeding, pregnant, lactation, bowel problem (defecation), as reported in recent studies [25]. Besides the above basic set of confounders, we also show the results adjusted for more potential confounders including dietary (dietary taste spicy, sweet, salty, oil, or light, high-sugar and high-fat diet habit, fruit and vegetable intake, and favors fat meat), exercise (exercise frequency, exercise intensity, and average time per exercise), drinking, smoking, and Bristol’s stool score.

2.15. Mathematical transformation for compositional data

We use mmvec [26,27] to re-rank and highlight the associations selected by average rank methods (see methods 2.17), which account for the composition effect. Gut relative abundance profile is a compositional data (sum to 1), “mmvec” (microbe-metabolite vectors) was a compositional technique that can handle multilomics datasets. It reports co-occurrence probabilities between microbes and metabolites by a two-layer neural network. Softmax transform enforces scale invariance to remove the composition effect. It has several users defined parameters. In our study, we set “-latent-dim 3” “-min-feature-count 10” “-learning-rate 1e-5” “-beta1 0.9” “-beta2 0.99.” Both the logloss and cv_rmse curves decaying and plateau is close to zero, which suggested a good model. Other parameter sets can achieve the same good fitting too, such as “-latent-dim 3” “-min-feature-count 10” “-learning-rate 1e-5” “-beta1 0.9” “-beta2 0.95” and “-latent-dim 5” “-min-feature-count 10” “-learning-rate 1e-5” “-beta1 0.85” “-beta2 0.9.” But the ranks will vary little across different parameter sets. The ranks vary by the log\((x+1)\) or tmm transform of metabolite profile in R package edgeR [28]. The best rank for a paired metabolite and microbe among different parameter sets and transform are reported in Supplementary Table 4b.

2.16. Benjamini-Hochberg multiple hypothesis testing adjustment

The multiple hypotheses testing Benjamini-Hochberg adjustment was performed for one source-target omics pair each time for Fig. 4. In Supplementary Table 2a, we show two versions of Benjamini-Hochberg adjustment results for Shannon and other types of omics data. Q-value was the BH adjustment within one omics each time. Q-value(all) was done overall on all omics.

2.17. Robust association network construction between any two omics data type, including fecal microbial MGSs

A rank average method [29] was used to combine the results of multiple inference methods to make a robust omics association network. We combined two nonlinear models, one-to-many random-forest and multiple inference methods to make a robust omics association network.

Step 1: Data preprocessing.

Dichotomous variables (such as gender) and unordered categorical variables (such as region) were recoding into dummy variables. Frequency items such as yogurt eating habits were assigned to integers. We kept microbial species with occurrence higher than 10 % for further analysis. For other type of omics data, we removed variables following these rules: (i) Near zero variance and (ii) With more than 30 % missing value. After filtering, we obtained 98 metabolites, 69 items in lifestyle questionnaire, 73 medical test parameters, 18 items in psychological questionnaire, 24 physical fitness test scores, and 2352 MLG/MGS. Missing values were filled with the median. Outliers were defined as outside of the 95 % quartiles and outlier samples were removed.

Step 2: Computation of associations using multiple inference methods.

For each factor in one omics, we performed regression using theRFCV function (R package RF) with default parameters based on all factors in another omics and calculated RFCV R [30]. Fivefold average variable importance was extracted for step 3. That is an average of five RF importance score for each fold. Partial Spearman’s correlation (Pearson function in R package ppcor) between factors from any two types of omics was also extracted. Potential confounders were considered as described above. We also show generalized linear model results from MaAslin [31] with default parameters after adjusting the above confounders.

Step 3: Robust networks construction.

To obtain the most robust and strongest association between factors from any two types of omics, in other words, to filter predictor factors and target factors, we performed two steps. First, the top 20 target factors with the highest RFCV R were selected as target factors. Then, to choose predictor factors for every selected target factor, we kept the predictor factors with top 30 average ranks, and retained the edges with partial Spearman’s correlation BH-adjusted p value < 0.05. The average rank was computed as the sum of the ranks across the RFCV importance and absolute partial Spearman rho. For example, metabolites as target and gut microbe as the source. We regressed gut microbes against the metabolites and compute the fivefold cross-validation predict power (RFCVR) and partial Spearman correlation for each metabolite. In all, 20 metabolites with the highest RFCVR were kept. For each of the 20 select metabolites such as VA, average ranks across RFCV, and partial Spearman were performed. Gut microbe biomarker for VA was found with average rank top 30th and passed the partial Spearman BH-adjusted p value < 0.05.

Step 4: Network visualization.

For each target factor, the top 5–10 average ranks source factor in each source omics type were selected as representative factors to make barplots using ggplot2 package. The heatmap package was used to plot the common representative factors that could be predicted to be the strongest by multiple omics data types (Fig. 2c). All the source-target factors pair RFCVR (a is source, b is target and c is source, a target) was boxplot (Fig. 2b) using ggplot2. The ComplexHeatmap package in R was used to plot omics triadic relation (Fig. 4). CytoScape was also used to visualize the global omics network (Fig. 2a).

3. Results

3.1. The 4D-SZ cohort and its smaller scale validation

The gut microbiome data are an important part of this multomic cohort, with some volunteers collecting samples as early as 2015 constituting the time dimension in “4D.” Here, for the initial 4D-SZ cohort (Fig. 1, Supplementary Tables 1a-d), 2183 fecal samples were collected during a physical examination in the spring of 2017 in Shenzhen, China. The gender ratio was close to 1:1 (1016 females and 1007 males). These Chinese volunteers were mostly young (Age: 29.6 ± 5.5, average ± stdev, and Supplementary Fig. 1) and lean (BMI: 21.7 ± 3.8 and average ± stdev). All the volunteers completed questionnaires beforehand through a cell phone application, with 56 entries from the lifestyle questionnaire (2183 samples) and 18 entries from the psychological questionnaire (689 samples). Those who were pregnant or sick were advised to join at a later date. As can be seen from the 72 items in the medical test data (body measurements and routine blood test, 2081 samples, and Supplementary Table 1b), the individuals in the cohort were mostly free of cardiometabolic diseases (e.g., fasting blood glucose: 5.2 ± 0.5 mmol/l and low-density lipoprotein cholesterol: 2.7 ± 0.6 mmol/l). Chronic medication use was also rare in this young cohort, unlike older cohorts from European countries (Supplementary Fig. 1). Other omics data, including 104 plasma metabolites (2158 samples) and 24 physical fitness data (1935 samples), were collected from the same individuals (Fig. 1 and Supplementary Tables 1b-d).

The validation cohort was designed in the same manner but organized at smaller scales at multiple locations in China, later in 2017. Besides the current location, these 1404 individuals also differed by hometown location as compared to the initial cohort, with 43.43 % instead of 58.79 % from Central China, i.e., Hubei and Hunan provinces (Supplementary Table 1b, Fig. 1). In all, 570 participants were reported as
females and 480 were reported as males. The age (29.5 ± 5.2) and BMI (21.5 ± 0.2) distribution was also narrow (Supplementary Fig. 1). Lifestyle questionnaire (1050 samples), psychological questionnaire (851 samples), medical test data (508 samples), and plasma metabolites data (1093 samples) were also collected, while physical fitness test results were only available from 176 samples.

### 3.2. A high-quality dataset of 3587 metagenomic shotgun-sequenced fecal samples

Fecal DNA from the initial 4D-SZ cohort was subjected to metagenomic shotgun sequencing, yielding 82.95 ± 24.26 million high-quality nonhuman reads per sample (Supplementary Table 1a). The reads were mapped to a comprehensive human gut microbiome reference gene catalog containing 9.9 million genes, reaching a mapping rate of 80.1 ± 4.9%, indicating the saturation of the gene content in the gut microbiome [1,16]. The genes in the reference catalog were clustered according to co-abundance into 1507 metagenomic species (MGSs, Pearson’s correlation coefficient) and 2981 metagenomic linkage groups (MLGs, Kendall’s taut) [15-17,22], to include both known and unknown microbes. The relative abundances of these species or strains (MGSs and MLGs) were available for each sample based on reads mapped to each gene.

MetaPhlAn2 was used as an alternative method for taxonomic assignment (Supplementary Table 1e), which was highly similar to the MLGs/MGSs results for bacteria (For the same species, the median of Pearson’s correlation between the relative abundances according to MGSs and MetaPhlAn2 was 0.9525, Supplementary Table 4). According to MetaPhlAn2, 53.94% of the samples contained at least one kind of virus, including phages or prophages (relative abundance 0.579% ± 0.539% and mean ± sd); 35.51% of the samples contained at least one kind of archaea (relative abundance 0.011% ± 0.104%). Fungi were detected at an even lower relative abundance (0.001% ± 0.008%) and lower occurrence.

Fecal samples from the 1404 individuals in the validation cohort were sequenced and profiled in the same manner (Supplementary Tables 1c), yielding 82.95 ± 24.26 million high-quality nonhuman reads per sample.

### 3.3. The gut microbiome weakly associated with other omics

To obtain an overall idea of the relationship between omics data, an inter-omics prediction value between omics data was calculated using a fivefold cross-validated random forest model (RFCV, Fig. 2a, Methods 2.17, and Supplementary figure 1i). Medical test data showed the highest global systematic association with other omics data. The accuracy of prediction from medical test data to physical fitness data, from metabolites to medical test data, and from medical test data to metabolites data, 75% quantile reached R = 0.461, 0.399, and 0.603, respectively (Spearman correlation of predicted value versus measured value and 6.12E-5), indicating the saturation of the gene content in the gut microbiome composition (RFCV = 0.770, 0.782, 0.555, 0.281, and 0.616). Among the metabolites data, serum testosterone, isoleucine, valine, 3-methyl histidine, and leucine can be predicted precisely by medical test data, and from medical test data to metabolites data, 75% quantile reached R = 0.461, 0.399, and 0.603, respectively (Spearman correlation of predicted value versus measured value and 6.12E-5), indicating the saturation of the gene content in the gut microbiome composition (RFCV = 0.770, 0.782, 0.555, 0.281, and 0.616). These results could be further validated by the Mantel test [24] (Supplementary Fig. 3a).

As our “other genome,” the gut microbiome has relatively weak global systematic association with other omics data. However, the gut microbiome could still be used to estimate other individual omics factors in this cohort according to the RFCV prediction model, which was trained by the initial cohort and tested on the validation cohort. Gut microbiomes have predictive power for medical test data, including serum creatinine, hematocrit, serum uric acid (RFCV R = 0.42, 0.40, and 0.40), and BMI (RFCV R = 0.28). We also observe that BMI could explain the gut microbial composition, even though the cumulative effect size was in single digits (Supplementary Tables 2b and 2c), consistent with previous reports [4,32]. The narrow distribution of BMI in this cohort (21.7 ± 3.8, Supplementary Table 1b) and perhaps the short exposure time for obesity-related comorbidities contributed to the small effect size of only 0.0015 for BMI (q-value = 0.014 and Supplementary Table 2b).

ABO blood group could also be weakly predicted by the fecal microbiome composition (RFCV R = 0.2 and Fig. 3), and specific differences include Lachnospiraceae bacterium 3 146FAA in blood type A (q = 6.12E-5), Ruminococcus torques in blood type B (q = 1.59E-2),
unnamed MGS209 in blood type AB ($q = 1.59E-2$), and Megasphaera micronuformis in blood type O ($q = 1.69E-2$).

The gut microbiome showed the greatest predictive power for metabolites in terms of a number of associated metabolites, such as plasma vitamins (vitamin A, folic acid, vitamin B5, vitamin D, and RFCV $R = 0.28, 0.22, 0.14, 0.14$), plasma hormones (testosterone, aldosterone, and RFCV $R = 0.36$ and $0.20$), trace elements (mercury, selenium, arsenic, and RFCV $R = 0.28, 0.18, 0.15$) and plasma amino acids (branched chain amino acids (BCAA), glutamic acid, tryptophan, tyrosine, histidine, alanine, and RFCV $R = 0.27, 0.21, 0.18, 0.30,$ and $0.24$) (Fig. 3). For the validation cohort with 1404 individuals, which differed by sampling location and hometown location as compared to the initial cohort (Supplementary Table 1a and 1b, and Fig. 1), the gut microbiome could also predict these plasma metabolites, with greater effects for mercury, cysteine, selenium, iron, and cobalt (Fig. 3, Supplementary Table 3, and RFCV $R = 0.43, 0.17, 0.44, 0.35,$ and $0.20$).

Fig. 2. Overview of the interrelationship between omics in the main cohort. a, Global association between omics datasets. Each arrow represents a 5-fold cross-validation random forest (RFCV) prediction. The direction of the arrow indicates the direction of prediction, and the source omics dataset used to predict the target dataset. The darkness and size of the arrow lines indicate 75% quantile of Spearman's correlation between the measured value and fivefold cross-validation RFCV predicted value (RFCV $R$). b, Detailed predictive power of source omics for each target omics. The tick label on x-axis represents target omics. The title on top represents source omics. Each node in the box is a target factor. The color of the node and box line indicates the target omics data. Y-axis represents the target factor RFCV $R$ predicted from source omics. c, Common representative factors that could be most strongly predicted by multiple omics data type. Y-axis tick label indicates source omics. The title is target omics. X-axis tick label represents common representative factors (target factors). The color in the heat map indicates the RFCV $R$ using the omics data on y-axis to predict each factor on x-axis. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)
3.4. Associations with circulating levels of vitamin A, trace elements, and amino acids

We next explored the metabolome-gut microbiome associations in detail. Vitamin A is central to a healthy immune system but is typically studied for its role in early development [33]. A recent mouse study reported modulation of retinol dehydrogenase 7 expression and dampened antimicrobial response in the gut by Clostridiales [34]. Consistently, we observed associations between Clostridia species (Clostridia MGS0123, MGS0560, MGS0558, Lachnospiraceae bacterium 1 456FAA, Lachnospiraceae bacterium 6163FAA, Lachnospiraceae bacterium 9143BFAA, Clostridium bolteae, Clostridium sp. AT4, and Clostridium sp. M62.1) and vitamin A in human adults with both Spearman's correlation and MaAsLin association (Fig. 4 and Supplementary Table 3a). Association between Lachnospiraceae bacterium 9143BFAA, C. bolteae, and vitamin A can also be validated in MetaPhlAn2 profile with GLM model or mmvec [26] model, which corrects for the compositional data (Supplementary Tables 3a and 4b). 5-methyltetrahydrofolate (5-methyl-THF, the active form of folic acid) exhibited a positive correlation with Eubacterium eligens (Fig. 4), a butyrate-producing bacterium that was relatively depleted in atherosclerotic cardiovascular disease [17]. 5-methyl-THF was also negatively associated with Doroa and Blautia species (Fig. 4, Supplementary Tables 3a and 4b), which have been implicated in obesity and could metabolize formate [35–37]. Alanine and to a lesser extent glutamic acid showed negative associations with low BMI-related bacteria such as Alistipes shahii, Bacteroides cellobiosilyticus, Bacteroides intestinalis, Ruminococcus lactis, and E. eligens [17] in this large cohort (Fig. 4 and Supplementary Table 4b), consistent with higher glutamic acid in individuals with obesity or insulin resistance [36,38].

Association between the microbiome and trace elements including mercury, selenium, and arsenic might be surprising (Fig. 4). Selenium-containing rice is commercially promoted as anti-cancer food, and we found that the association pattern largely followed arsenic, consistent with these two trace elements' similar function in anaerobic respiration [39]. Selenium and mercury also correlated with disease-associated species such as C. bolteae and Ruminococcus gnavus in the gut microbiome (Supplementary Tables 3a and 4b). Plasma levels of phosphoserine, argininosuccinic acid, homocitrulline, serine, and glycine appeared as another cluster with similar associations with the fecal microbiome (Fig. 4). Phosphoserine and argininosuccinic acid were negatively associated with Bacteroides coprphilus (Supplementary Table 3a), a prevalent but not very abundant species from the Bacteroides genus. MGs from Faecalibacterium prausnitzii (Supplementary Figs. 4 and 5), a bacterium reported to produce butyrate...
and metabolize arsenic [40] was positively associated with L-homocitrulline and phosphoserine and negatively associated with vitamin A and mercury (Fig. 4 and Supplementary Table 3).

The 25-hydroxyvitamin D negatively correlated with the fungi *Saccharomyces cerevisiae* (MetaPhlAn2, Supplementary Table 4). Taken together, associations between the fecal metagenomic shotgun data and the metabolome data offer a number of hints for further investigation.

### 3.5. The gut microbiome contained information for defecation, hormone, and gender

Sex (females 1,016, males 1,007, and initial cohort) was one of the most significant factors to diverge gut microbiome composition (Fig. 3). *E. dolichum* and *Blautia wexlerae* were significantly more abundant in male than in female subjects, after adjusting for age, BMI, medication, and dietary supplements by partial Spearman’s correlation or generalized linear model (GLM) (Supplementary Table 3b and Supplementary Table 4c).

*Fusobacterium mortiferum*, which was positively associated with testosterone (Supplementary Table 3a), was only enriched in male subjects before but not after the statistical adjustments (Supplementary Tables 3b and 4c). Compared to male subjects, the fecal microbiome of female subjects showed a higher α-diversity (Supplementary Table 2a and Supplementary Fig. 4b). *Bifidobacterium longum*, *B. bifidum*, *B. catenulatum*, and *B. pseudocatenulatum* as well as potential oral or vaginal bacteria, such as *Streptococcus parasanguinis*, *Prevotella bivia*, and *Gardnerella vaginalis*, were all significantly enriched in female subjects (Supplementary Table 3b). Another fecal bacteria better known in the vaginal microbiota, *Lactobacillus crispatus*, was positively associated with hydroxyl progesterone, testosterone, and aldosterone (Supplementary Table 4b) consistent with a parallel study on the vaginocervical microbiome for this cohort [41].

Aldosterone, one of the major adrenal gland mineralocorticoid, was positively correlated with bacteria implicated in cardiometabolic health, such as *B. intestinalis*, *B. cellulosilyticus*, *B. stercorirosoris*, and *E. eligens* (Fig. 4) [17]. *E. eligens* and *Ruminococcus lactaris* scaled negatively with a self-reported preference for a salty diet, in contrast to *Blautia obeum*, which positively correlated with a salty diet (Supplementary Table 3b), mirroring the loss of commensal bacteria in mice on a high salt diet [42].

Self-reported defecation frequency was also among the strongest factors that were predicted by the gut microbiome composition (Fig. 3). Gut microbial functional potential for secondary bile acids strongly associated with self-reported defecation frequency, which was better validated than associations with sex hormones (Supplementary Fig. 3b), suggesting that these are stable patterns. These results extend previous findings regarding Bristol’s stool score (BSS), and potential sex differences in the microbiome (Supplementary Fig. 1) [4,32,43], down to species and bile acid gene level. In addition, the methanogenic archaea, *Methanobrevibacter smithii*, was associated with less frequent defecation; plasma level of vitamin B5, recommended for people with constipation, interestingly showed a negative correlation with the relative abundance of fecal *M. smithii* (MetaPhlAn2, Supplementary Table 4).

### 4. Discussion

In summary, our transomic investigation of several thousand volunteers established an unprecedented reference dataset for the human gut microbiome in the Asian population. A comprehensive reference dataset is essential to understand the microbiome background, particularly in pandemics. Judging from the associations, it appears as though a number of factors in circulation crosstalk with the gut microbiome, and then are reflected in the blood, brain, and fitness tests. Levels of trace elements, such as mercury, arsenic, and selenium, as important cofactors for bacteria respiration and other functions [38], should be measured even in uncontaminated regions and individuals showing normal levels of these
elements. Although rice is often studied for such contaminants, exposure can come from other food, drink, air, and soil sources [9,44]. Mice fed with arsenic (As) or cadmium (Cd) in drinking water showed changes in the gut microbiota [45]. Our results suggest that the metabolism of trace elements in commensal microbes might help determine their levels in the blood and influence immune functions. Cohort design in the future needs to take these environmental factors into consideration (Supplementary Fig. 2), and move beyond traditionally studied factors for the gut microbiome.

Here, the validation cohort was also mostly young adults in urban areas of China, and future international collaborations would be necessary to truly establish a baseline for the gut microbiome as well as other omics. The issue of compositional data (i.e., spurious negative correlations between taxa that are constrained by the sum of 1) is more challenging for low-diversity samples such as the vaginal microbiome or phyla-level claims from 16S rRNA gene amplicon data. Here, we only report associations between different omics, i.e., fecal microbiome vs. blood metabolites and fecal microbiome vs. physical fitness test results, which are not constrained by the sum of 1.

We have tentatively identified gut bacteria associated with each ABO blood type. A larger proportion of blood type A in Europeans than in East Asians might help explain the greater abundance of Lactobacillus bacterium [16,46]. Blood type B is more prevalent in northern Chinese, and the blood type B-enriched mucin-degrading bacterium R. torques has recently been reported to show an association with blood glucose [47] and was also associated with ulcerative colitis [48], and with a loose microbiota [49]. Genetic studies of the gut microbiota have not yet reported genome-wide significant associations with the ABO blood type genes themselves [50–53], while multiple studies have reported the impact of FUT2 secretor/nonsecretor status on gut microbiota composition [54–56]. Tentative associations here are yet to be matched with in vitro studies with the glycans [57,58].

Differences in gut microbiome composition between sexes and greater microbial diversity in female subjects have recently been reported in the LifeLines Deep cohort [32] and a Han Chinese cohort [59], yet the gut microbiome in female subjects was influenced by oral contraceptives, ovariectomy as well as antibiotics for vaginal or pelvic infections [43]. Males of Hadza hunter-gatherers showed differences in gut microbiota as compared to females [60], including higher Eubacterium and Blautia in men, which were recapitulated in our Chinese cohort (E. dolichum, B. wexlerae, and Supplementary Table 3b). Interestingly, E. dolichum associated with a dietary structure of more meat instead of fruit and vegetables, while B. wexlerae scaled negatively with milk consumption (Supplementary Table 3b). The prevalence of the Prevotellaceae family in our cohort is 16.0 ± 23.4%, which is comparable to the Hadza hunter-gather cohort (mean 29.8%, versus 7.6% in Italy) [60,61]. The evolutionary implications remain unclear.

We also find intriguing that decarboxylases appear generally important for bacterial stress response in the microbiome, i.e., to maintain a balanced pH for themselves. The top one for gut microbes may be glutamate decarboxylase (produces GABA (γ-aminobutyric acid) from glutamate), while histidine decarboxylase in the female reproductive tract might contribute to menstrual pains [41]. In addition, recent studies identified tyrosine decarboxylases in gut microbes that could digest the drug levodopa used to treat Parkinson’s disease [62,63].

Thus, this study provides a reference for the young gut microbiome with physical fitness test and questionnaire data (Supplementary Fig. 1), revealing interrelationships with other omics data such as trace elements and hormones that have so far not been included in other study designs (Supplementary Fig. 2). Even though we have tried to adjust for the known confounders, interventional as well as mechanical studies will be needed to confirm the association or causation between omics data.

Credit author statement


Data and materials availability

Metagenomic sequencing data for all samples have been deposited to the CNSA (https://db.cngb.org/cnsa/) of (CNGB) database under the accession code CNP0000426 [64,65] and CNP0000289 [64,65]. The main R code for calculating the RFCV R is available at https://github.com/jiezhuomicomics.

Declaration of competing interest

The authors declare no competing financial interest.

Acknowledgments

This study was funded by the Guangdong Key Areas R&D Projects, China (No. 2019B020229003). The authors are very grateful to colleagues at BGI-Shenzhen and China National Genebank (CNGB), Shenzhen, for sample collection, DNA extraction, library construction, sequencing, and discussions. We thank Dr. Qiang Sun (University of Toronto), our colleagues Chen Chen and Yanmei Ju, Jinghua Wu, and Zhe Zhang for their helpful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.medmic.2021.100039.

References

[24] Mantel N. The detection of disease clustering and a generalized regression
[39] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[38] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[39] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[36] Smyth GK. edgeR: a Bioconductor package for differential expression analysis of
[39] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[36] Smyth GK. edgeR: a Bioconductor package for differential expression analysis of
[39] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[36] Smyth GK. edgeR: a Bioconductor package for differential expression analysis of
[39] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[36] Smyth GK. edgeR: a Bioconductor package for differential expression analysis of
[39] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[36] Smyth GK. edgeR: a Bioconductor package for differential expression analysis of
[39] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[36] Smyth GK. edgeR: a Bioconductor package for differential expression analysis of
[39] Stolz JF, Oremland RS. Bacterial respiration of arsenic and selenium. FEMS
[36] Smyth GK. edgeR: a Bioconductor package for differential expression analysis of