



Published in final edited form as:

*Metabolomics*. ; 16(10): 103. doi:10.1007/s11306-020-01724-9.

## Urbanization in China is associated with pronounced perturbation of plasma metabolites

Yiqing Wang<sup>a,\*</sup>, Wei Sha<sup>b,c,\*</sup>, Huijun Wang<sup>d</sup>, Annie Green Howard<sup>e,f</sup>, Matthew C. B. Tsilimigras<sup>a,e,g</sup>, Jiguo Zhang<sup>d</sup>, Chang Su<sup>d</sup>, Zhihong Wang<sup>d</sup>, Bing Zhang<sup>d,†</sup>, Anthony A. Fodor<sup>b</sup>, Penny Gordon-Larsen<sup>a,e,†</sup>

<sup>a</sup>Department of Nutrition, Gillings School of Global Public Health & School of Medicine, University of North Carolina at Chapel Hill (UNC-Chapel Hill), NC;

<sup>b</sup>Department of Bioinformatics and Genomics, University of North Carolina at Charlotte (UNCC), NC;

<sup>c</sup>Department of Cancer Biostatistics, Levine Cancer Institute, Atrium Health, Charlotte, NC;

<sup>d</sup>National Institute for Nutrition and Health, Chinese Center for Disease Control and Prevention, Beijing, China;

<sup>e</sup>Carolina Population Center, UNC-Chapel Hill, NC;

<sup>f</sup>Department of Biostatistics, Gillings School of Global Public Health, UNC-Chapel Hill, NC;

<sup>g</sup>Department of Epidemiology, Gillings School of Global Public Health, UNC-Chapel Hill, NC;

### Abstract

---

Please Address Correspondence and Reprint Requests To: Penny Gordon-Larsen, PhD, University of North Carolina at Chapel Hill; Carolina Population Center, 123 West Franklin Street, Chapel Hill, NC 27516, Phone: (W) 919-962-6110; Fax: 919-966-9156, [pglarsen@unc.edu](mailto:pglarsen@unc.edu).

Authors' contributions

HW, JZ, CS, ZW, and BZ: performed data collection. YW and WS: designed study, analyzed data, and wrote the paper. AGH, AAF and MCBT: supervised data analysis and provided intellectual content to revise the manuscript. PG-L: designed study and provided intellectual content to revise the manuscript. All authors read and approved the manuscript.

\*These authors contributed equally;

†These authors contributed equally.

Conflicts of interest/Competing interests

The authors declare that they have no conflict of interest.

Compliance with Ethical Requirements

The study met the standards for the ethical treatment of participants and was approved by the Institutional Review Boards of the University of North Carolina at Chapel Hill, Chinese Center for Disease Control and Prevention, and the National Institute for Nutrition and Health.

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Availability of data and material

The metadata associated to the samples, including community-, household-, and individual-level data like urbanization index, household income, education, and physical activity, is publicly available at the China Health and Nutrition Survey (CHNS) website: <https://www.cpc.unc.edu/projects/china>. The metabolomics data analyzed during the current study are currently not publicly available, but are pending release to a public repository with an embargo in the future. Metabolomics data are available to the journal editors (in de-identified form) for replication and collaborators by permission from the corresponding author.

Code availability

The code used for the current study are available upon reasonable request.

**INTRODUCTION:** Urbanization is associated with major changes in environmental and lifestyle exposures that may influence metabolic signatures.

**OBJECTIVES:** We investigated cross-sectional urban and rural differences in plasma metabolome analyzed by liquid chromatography/mass spectrometry platform in 500 Chinese adults aged 25–68 years from two neighboring southern Chinese provinces.

**METHODS:** We first examined the overall metabolome differences by urban and rural residential location, using Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) and random forest classification. We then tested the association between urbanization status and individual metabolites using a linear regression adjusting for age, sex, and province and conducted pathway analysis (Fisher's exact test) to identify metabolic pathways differed by urbanization status.

**RESULTS:** We observed distinct overall metabolome by urbanization status in OPLS-DA and random forest classification. Using linear regression, out of a total of 1108 unique metabolite features identified in this sample, we found 266 metabolites that differed by urbanization status (positive false discovery rate-adjusted p-value, q-value<0.05). For example, the following metabolites were positively associated with urbanization status: caffeine metabolites from xanthine metabolism, hazardous pollutants like 4-hydroxychlorothalonil and perfluorooctanesulfonate, and metabolites implicated in cardiometabolic diseases, such as branched-chain amino acids. In pathway analysis, we found that xanthine metabolism pathways differed by urbanization status (q-value=1.64E-04).

**CONCLUSION:** We detected profound differences in host metabolites by urbanization status. Urban residents were characterized by metabolites signaling caffeine metabolism and toxic pollutants and metabolites on known pathways to cardiometabolic disease risks, compared to their rural counterparts. Our findings highlight the importance of considering urbanization in metabolomics analysis.

### Keywords

Urbanization; Plasma metabolomics; Pollution; China

## INTRODUCTION

Urbanization results in many changes in environments and lifestyles including a shift from traditional to Western diet, characterized by increasing intakes of animal-source foods, fats, added sugar, and refined carbohydrates (Lin et al., 2017, Popkin et al., 2012). Other lifestyle health behaviors also change along with urbanization, such as a decline in physical activity (Ng et al., 2014). In addition, environments change with urbanization, including increases in chemical toxicant exposures like perfluorooctanesulfonate (PFOS) (Chen et al., 2009). However, few studies have examined how urbanization may relate to differences in circulating metabolites between urban versus rural residents.

Recent advancements in metabolomics profiling enable examination of potential molecular mechanisms of lifestyle-related diseases that are becoming more common in urbanizing countries. The host circulating metabolome, a collection of all biological processes reflecting genetic and exogenous influences, have been suggested to be markedly altered by

the gut microbiome (Zhernakova et al., 2018), diseases (Menni et al., 2015), drug intake (Hiltunen et al., 2017), in addition to environmental exposures (van Veldhoven et al., 2019), diet (Shin et al., 2019), and physical activity (Ding et al., 2019). For example, the Oxford Street II study showed that short-term exposure to traffic-related air pollution in London was associated with various metabolic features in serum, including decreased levels of acyl-carnitines (van Veldhoven et al., 2019), which are known to regulate energy metabolism in cardiac mitochondria (Makrecka et al., 2014). Another randomized crossover trial suggested that serum ketone bodies were elevated in a typical American diet (Shin et al., 2019), while a traditional Korea diet characterized by low intakes of animal-source foods and high intakes of vegetables and whole grains was associated with decreased levels of branched-chain amino acids (BCAAs), which have been linked to diabetes and obesity (Yoneshiro et al., 2019, Newgard et al., 2009) (Zhou et al., 2019). Yet, these studies tend to focus only on a single domain of urbanization-related exposures, rather than the total set of exposures captured by a simple urban versus rural comparison, particularly if comparing urban and rural areas of different continents.

Given the lack of data on rural and urban residents living in the same geographic areas to control for regional and cultural influences relative to urbanization influences, we used plasma metabolomics data from 500 adult participants of the 2015 China Health and Nutrition Survey (CHNS) to examine the differences in host metabolites between urban (n=240) and rural (n=260) residential locations. These adults were from two neighboring southern provinces, Hunan and Guizhou, that were similar in geography and dietary patterns. We first examined urban versus rural differences in the overall metabolome. Then, we tested individual metabolites differences by urbanization status and conducted pathway analysis. In secondary analysis, we examined the associations between specific urbanization-related lifestyle and behavioral factors with urbanization-associated metabolites identified from our analyses of individual metabolites.

## METHODS

### Study population

The CHNS is a household-based longitudinal study across 12 provinces and three megacities in China and is designed to capture urban and rural differences in socioeconomics, public resources, health behaviors, and health status. The original subjects in 1989 were selected using a stratified, multistage random cluster design as described previously (Popkin et al., 2009). For the current cross-sectional analysis, we used year 2015 data from a subset of adults (n=500) aged 25–68 years from two neighboring Southern provinces, Hunan and Guizhou. The study met the standards for the ethical treatment of participants and was approved by the Institutional Review Boards of the University of North Carolina at Chapel Hill, Chinese Center for Disease Control and Prevention, and the National Institute for Nutrition and Health.

### Plasma metabolomic profiling

Fasting blood was collected via venipuncture with ethylenediamine tetraacetic acid (EDTA) as an anticoagulant, by certified clinicians from the local China Center for Disease Control

and Prevention (CCDC) with extensive training and experience in processing clinical fasting blood samples. Blood samples were immediately refrigerated at  $-2-8^{\circ}\text{C}$ , and within three hours, samples were transported in refrigerator to laboratories and centrifuged to separate plasma, which were stored at  $-80^{\circ}\text{C}$  until processing. All sites followed the same standardized protocol for the collection, processing, and storage of blood samples with strict quality control. Our laboratories were accredited in College of American Pathologists (CAP) laboratory accreditation program and International Organization for Standardization (ISO) 15189 program.

The global metabolomics analysis was performed using an integrated, ultrahigh performance liquid Chromatography-tandem mass spectrometry (UPLC-MS/MS) at Metabolon's partner campus in China. Each sample was accessioned into the Metabolon Laboratory Information Management System (LIMS) system and assigned a unique identifier, which was used to track all sample handling, tasks, results, etc. Samples were processed by an automated MicroLab STAR system (Hamilton Company), with several recovery standards and controls added prior to the first step in extraction as technical replicates for quality control (QC), which were DL-2-fluorophenylglycine, tridecanoic acid, d6-cholesterol and 4-chlorophenylalanine (Evans et al., 2014). Proteins were precipitated with methanol under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000) followed by centrifugation to remove protein, dissociate small molecules, and recover chemically diverse metabolites. The resulting extract was divided into five fractions: two samples for analysis by two separate reverse phase (RP)/UPLC-MS/MS with positive ion mode electrospray ionization (ESI); one for analysis by RP/UPLC-MS/MS with negative ion mode ESI; one for analysis by Hydrophilic Interaction Liquid Chromatography (HILIC)/UPLC-MS/MS with negative ion mode ESI; and one for backup.

A Waters ACQUITY UPLC (Waters, Milford, MA) and a Thermo Scientific Q-Exactive high-resolution/accurate MS (ThermoFisher, Waltham, MA), interfaced with heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer (35,000 mass resolution) were used in all methods. The extract was dried then reconstituted in solvents compatible to each method. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. Specifically, one aliquot of each sample extract was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds: the extract was gradient eluted from a C18 column (Waters UPLC BEH C18- $2.1\times 100$  mm,  $1.7\ \mu\text{m}$ ) using water and methanol, containing 0.05% perfluoropentanoic acid (PFP) and 0.1% formic acid (FA). A second aliquot was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophobic compounds: the extract was gradient eluted from the same afore-mentioned C18 column using methanol, acetonitrile, water, 0.05% PFP and 0.01% FA and was operated at an overall higher organic content. A third aliquot was analyzed using basic negative ion optimized conditions using a separate dedicated C18 column. The basic extracts were gradient eluted from the column using methanol and water, with 6.5mM Ammonium Bicarbonate at pH 8. The fourth aliquot was analyzed via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide  $2.1\times 150$  mm,  $1.7\ \mu\text{m}$ ) using a gradient consisting of water and acetonitrile with 10mM Ammonium Formate, pH

10.8. The MS analysis alternated between MS and data-dependent MS<sup>n</sup> scans using dynamic exclusion, which ranged 70–1000 m/z.

For quality assurance, the following types of controls were analyzed in concert with the experimental samples: a pooled sample generated by taking a small portion of each experimental sample (i.e., technical replicate), a pool of well-characterized human plasma maintained by Metabolon, extracted water samples (i.e., process blanks), aliquot of solvents used in extraction (i.e., solvent blanks), and a cocktail of QC standards that were carefully chosen not to interfere with the measurement of endogenous compounds were spiked into every analyzed sample, allowed instrument performance monitoring and aided chromatographic alignment. Experimental samples were randomized across the platform run with QC samples spaced evenly among the injections. Instrument (6%) and overall process variabilities (13%) were determined by calculating the median relative standard deviation (RSD) for the standards added to each sample prior to injection into the mass spectrometers and for all endogenous metabolites present in all of the pooled matrix samples.

Raw data was extracted, peak-identified, and QC processed using Metabolon's hardware and software built under Microsoft's .NET framework. Chemicals were identified by comparing to purified standards or recurrent unknown entities in Metabolon's library, which maintained based on authenticated standards containing mass-to-charge ratio (m/z), chromatographic data, MS/MS data, and retention time/index (RI). RI of each compound was based on its elution relationship (assuming a linear fit) with two surrounding standards, which were isotopically labeled metabolites given a fixed RI value (Evans et al., 2009). Metabolites were identified based on three criteria: RI within a narrow RI window of the proposed identification, accurate mass match to the library (+/- 10 ppm), and the MS/MS forward and reverse scores between the experimental data and authentic standards. The MS/MS scores are based on a comparison of the ions present in the experimental spectrum to the ions present in the library spectrum. While there may be similarities between these molecules based on one of these factors, the use of all three data points can be utilized to distinguish and differentiate biochemicals. More than 3300 commercially available purified standard compounds have been acquired and registered into LIMS for analysis on all platforms for determination of their analytical characteristics. Additional mass spectral entries have been created for structurally unnamed biochemicals, which have been identified by virtue of their recurrent nature (both chromatographic and mass spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis.

Metabolon data analysts use proprietary visualization and interpretation software to confirm the consistency of peak identification among the various samples and to remove system artifacts, mis-assignments, and background noise. Library matches for each compound were checked for each sample and corrected if necessary. Peaks were quantified using area-under-the-curve. A total of 1108 unique metabolite features were detected and quantified in our sample, among which 725 metabolites were at Level 1 identification according to the Metabolomics Standards Initiative (Sumner et al., 2007); 165 metabolites were at Level 2 and labeled as "Biochemical Name\*"; 17 metabolites were at Level 3 and labeled as "Biochemical Name\*\*"; 200 metabolites were at Level 4 and each was assigned a unique

number after “X -”. Each compound was corrected in run-day blocks by registering the medians to equal one with values below detection limit imputed by the minimum detected quantity.

### Urbanization assessment

To account for potential misclassification of a dichotomous measure of urban-rural environments based solely on population density, we employed a validated urbanization index that encompasses 12 dimensions of urbanized environments that go beyond the typical standards of population size and/or density (Jones-Smith and Popkin, 2010). Specifically, our urbanization index distinguishes urban characteristics on a continuous scale (0–120 points) based on the following 12 measures with each allotted a maximum of 10 points: population density, economic activity, traditional markets, modern markets, transportation infrastructure, sanitation, communications, housing, education, diversity, health infrastructure, and social services, which were measured using standardized community and household questionnaires. This urbanization index has been shown to relate to health outcomes, reflecting its utility in studies of urbanization and health outcomes (Jones-Smith and Popkin, 2010, Inoue et al., 2018). Using this complex measure of urbanization, we dichotomized urbanization index by median to define urban (76.1–99.6) and rural areas (39.2–73.1).

### Urbanization-related factors

Sociodemographic and behavioral factors were measured using questionnaires during 5-day household visits. We dichotomized educational attainment by high school completion. We calculated per capita household income by dividing the total gross household income by the number of household members. Physical activity in metabolic equivalent of tasks (METs) per week were calculated using 7-day recalls of all domestic, occupational, transportation, and leisure activities using the Compendium of Physical Activity (Ng and Popkin, 2012). Dietary intakes were assessed using three consecutive 24-hour recalls by trained interviewers and household food inventories during the same 3-day period. Three-day average nutrient intakes were estimated using a Chinese food composition table (Yang, 2005), with total energy intake validated by doubly labeled water (Pearson correlation coefficient men: 0.56; women: 0.60) (Yao et al., 2003). We calculated percent energy (% kcal) from animal-source foods, a strong indicator of Westernized diet in China (Popkin and Du, 2003). We estimated diet diversity using the number of food groups consumed, including staple grains, fruits, vegetables, nuts, eggs, meat, seafood, and dairy products.

### Statistical analysis

We used chi-squared tests and Wilcoxon rank sum tests to compare categorical and continuous characteristics with non-normal distributions between urban and rural participants, respectively.

To provide insights into the urban versus rural differences in the overall metabolome, we first performed a supervised multivariate analysis, Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) using the SIMCA software (Version 16, Umetrics, Umeå, Sweden). We used a 7-round cross-validation to compute the diagnostic Q<sup>2</sup>Y value,

which is a measure of model predictive ability. To prevent overfitting, we used permutation-based validation, during which  $Q^2Y$  of the model using un-permuted data was compared with the  $Q^2Y$  of each model built using each of the 999 permuted data sets. Our model passed this permutation-based validation (Supplementary Fig. 1), as  $Q^2Y$  of the model using un-permuted data was found to be better than any of the models built using the permuted data sets. Additionally, we performed principal component analysis (PCA) as a quality assessment tool for OPLS-DA. In OPLS-DA, we used the Variable Importance in Projection (VIP) score, which reflects the variability of urbanization status explained by each metabolite and provides loading weights for each metabolite (Mehmood et al., 2012), to assess the contribution of each metabolite to the urban-rural separation.

Next, we used random forest classification (10000 trees) with hold-out validation in JMP Pro (version 13, SAS institute, Cary, NC) to determine whether urban and rural participants could be discriminated based on overall metabolome data. We randomly split data into training (70%), validation (15%), and test (15%) datasets. The distribution of urban versus rural participants were roughly balanced across training (e.g., 46.6%), validation (51.9%), and test (50.0%) datasets, which was similar to the full sample (48%). We evaluated the classification performance based on sensitivity and specificity using receiver operating characteristic (ROC) curve and metabolite importance based on training dataset using  $G^2$ , the likelihood-ratio of chi-square when a given metabolite was added to the tree.

To identify metabolites that differed by urbanization status, we used a linear regression model adjusted for age, sex, and province to examine the associations between urbanization status and individual metabolites relative abundance, which was  $\log_2$  transformed for normality assumption. We exponentiated regression coefficients with base two to calculate the urban versus rural ratio and adjusted p-values for multiple hypothesis testing for 890 metabolites at the Level 1 and Level 2 assignment using positive false discovery rate (pFDR, q value) (Storey et al., 2004, Storey, 2002).

Additionally, we performed pathway analysis on the Metabolon portal (Metabolon Inc., Durham, NC) to investigate which metabolic pathways demonstrated statistically significant differences between urban and rural participants. We calculated enrichment score as  $(k/m)/[(n-k)/(N-m)]$  for each pathway that reflect the degree to which a given pathway was overrepresented in either urban or rural participants, where k and m are numbers of significant metabolites ( $q\text{-value} < 0.05$  in linear regression) and detected metabolites in the given pathway, respectively, and n and N are numbers of significant metabolites ( $q\text{-value} < 0.05$  in linear regression) and detected metabolites in all pathways, respectively. We used Fisher's exact test to estimate p-value for each pathway, which was adjusted by pFDR. We removed 218 Level 3 and Level 4 metabolite features from pathway analysis, giving a total of 890 metabolites from 112 metabolic pathways.

In secondary analyses, we assessed the associations between the following six urbanization-related factors: education, income, physical activity, energy intake, animal-source foods (%kcal), and diet diversity, with  $\log_2$  transformed relative abundance of urbanization-associated metabolites identified in linear regression models ( $q\text{-value} < 0.05$ ) adjusting for age, sex, and province. We exponentiated regression coefficients with base two and thus an

exponentiated coefficient  $>1$  indicates a positive association and vice versa. We adjusted p-values using pFDR (Storey et al., 2004, Storey, 2002). We used SAS (Version 9.4, SAS Institute Inc., Cary, NC) and R 3.6.0 (<http://www.r-project.org>) to perform statistical tests.

## RESULTS

Our sample was similar in age and sex across rural and urban areas (Table 1). We found that compared to rural participants, urban participants had higher educational attainment, income, and diet diversity, but lower physical activity. Total energy intake and animal-source food consumption did not differ by urbanization status.

Using OPLS-DA and random forest classification, we assessed the discrimination between urban and rural participants based on their overall metabolome (Fig. 1). We observed a good separation (model explained variation:  $R^2Y=58.7\%$ , model predictive ability:  $Q^2Y=35.5\%$ ) of the overall metabolome by urbanization status on the horizontal OPLS-DA axis (Fig. 1a). The PCA biplot also showed a good separation between urban and rural samples (Supplementary Fig. 2). In random forest classification to determine how well metabolites on their own could distinguish urban and rural participants (Fig. 1b), the high values of areas under the ROC curve (AUC) for training (1.00), validation (0.91), and test (0.88) datasets indicated reliable and powerful models with high sensitivity and specificity. To further aid in interpreting Fig. 1, we show metabolites in order of their contributions to urban and rural separation in OPLS-DA (VIP score) and random forest classification ( $G^2$ ) in Supplementary Tables 1 and 2, respectively. A metabolite of fungicide chlorothalonil, 4-hydroxychlorothalonil ( $G^2=15.83$ ) had the highest contribution to the random forest classification, followed by theobromine (xanthine metabolism,  $G^2=2.86$ ), gamma-tocopherol/beta-tocopherol (tocopherol metabolism,  $G^2=2.35$ ), cis-3,4-methyleneheptanoyl carnitine (medium-chain acyl-carnitine,  $G^2=1.69$ ), perfluorooctanesulfonate (PFOS, xenobiotic chemical,  $G^2=1.20$ ), and 1-methylxanthine (xanthine metabolism,  $G^2=1.11$ ).

To identify the association between urbanization status and individual metabolites, we used a linear regression adjusted for age, sex, and province and found that urbanization status was associated with 266 classified metabolites at  $q\text{-value}<0.05$  (Supplementary Table 3), including 112 lipids, 58 amino acids, and 47 xenobiotics. Among these urbanization-associated metabolites, 84 and 30 metabolites were the top contributors in OPLS-DA (VIP 1.5) and random forest classification ( $G^2 > 0.5$ ), respectively. Urbanization status had the strongest association with 4-hydroxychlorothalonil ( $q\text{-value}=2.03E-23$ ), a known pollutant, with 2.91 (95% confidence interval: 2.41, 3.51) times the relative abundance of 4-hydroxychlorothalonil in urban versus rural participants. Additionally, urban status was associated with 1.29 (1.1, 1.5) times as much PFOS as rural status in units of relative abundance ( $q\text{-value}=5.22E-03$ ). Among the 15 metabolites at  $q\text{-value}<1.0E-05$  (including three partially characterized metabolites), the relative abundance of six metabolites from the xanthine metabolism pathway (theobromine, theophylline, paraxanthine, 1,7-dimethylurate, 1-methylxanthine, caffeine) was more than doubled in urban versus rural participants. These results are summarized in a volcano plot in Fig. 2. The boxplots show the unadjusted differences in the distributions of relative abundance by urbanization status (Supplementary Fig. 1), for those 15 metabolites ( $q\text{-value}<1.0E-05$ ) where we found evidence of differences



by urban versus rural status in the linear regression models adjusted for age, sex, and province.

In pathway analysis, we assessed the degree of urban-rural difference for each metabolic pathway measured by enrichment score and identified the pathways that differed by urbanization status using Fisher exact test. We found that xanthine metabolic pathway had the highest enrichment score (3.45) and statistically significant difference between urban and rural participants ( $q$ -value=1.64E-04, Table 2, pathways ordered by  $q$ -value). While the other pathways were not statistically significant after pFDR adjustment, some were nominally significant at  $p$ -value<0.05, including sphingomyelins, glutathione, dipeptide, acyl-glycine, and tocopherol metabolic pathways. In Fig. 3, we show that all of the identified metabolites in the xanthine metabolism pathways were derived from caffeine.

In secondary analysis to examine the association between key urbanization-related lifestyle and behavior factors with urbanization-associated metabolites, we used a linear regression adjusting for age, sex, and province and found clear differences in pathways and direction of associations across four factors (education, physical activity, energy intake, and diet diversity) (Fig. 4, Supplementary Table 4–7). Specifically, we found that high school education, physical activity, total energy intake, and diet diversity were associated with 107, 103, 11, and 100 of the 266 metabolites identified by the linear regression at  $q$ -value<0.05, respectively, whereas per capita household income and animal-source food consumption were not associated with any of these metabolites. All of these factors were associated with HXGXA (polypeptide), leucylglutamine (dipeptide), and PFOS. Metabolites positively associated with urbanization status (i.e. urban versus rural ratio >1 and  $q$ -value <0.05) tended to be positively associated with education, energy intake, and diet diversity, but negatively associated with physical activity, including xanthine metabolites, cysteinylglycine disulfide from glutathione metabolism, and two known pollutants, 4-hydroxychlorothalonil and PFOS.

## DISCUSSION

Using plasma global metabolomics analysis in 500 Chinese adults from two neighboring provinces, we found profound urban and rural differences in overall host metabolome and in individual metabolites. Specifically, we observed a separation of the overall metabolome by urbanization status in OPLS-DA and good discrimination (AUC=0.88 in test datasets) of urban and rural participants based on their overall metabolome in random forest classification. Additionally, we identified 266 metabolites that differed across rural and urban areas using linear regression, including xenobiotics from dietary sources and pollutions, such as caffeine metabolites, 4-hydroxychlorothalonil, and PFOS, as well as metabolites known to be involved in etiologies of cardiometabolic diseases, such as BCAAs (Newgard et al., 2009) and fibrinogen peptide A (Eisenberg et al., 1985), suggesting potential differences in environmental exposures and health status between urban and rural adults.

Among all detected metabolites, 4-hydroxychlorothalonil had the strongest statistical association with urbanization status in the linear models (lowest  $q$ -value and second largest

urban versus rural ratio) and the greatest contribution to distinguishing urban and rural participants using random forest classification. The almost tripled relative abundance of 4-hydroxychlorothalonil in urban versus rural participants on average was concerning because 4-hydroxychlorothalonil is a major metabolite of chlorothalonil, which is mainly used as a broad-spectrum fungicide and a probable human carcinogen (Rossi, 1999) that has been banned by both European Union and Switzerland since 2019. Several unpublished studies using mice and rat models reported by the United States Environmental Protection Agency (EPA) suggested that long-term dietary exposure to chlorothalonil may relate to greater incidence of renal and stomach tumors (Rossi, 1999), and one study using amphibians showed that environmental chlorothalonil, at a concentration to which humans are commonly exposed, may relate to higher mortality, elevated corticosterone levels, and damaged immune cells (McMahon et al., 2011). A few unpublished studies reported by the EPA also suggested that 4-hydroxychlorothalonil may relate to renal tubular degeneration in dogs and reduced pup body weight in rats (Rossi, 1999), and one study using zebrafish showed that 4-hydroxychlorothalonil may have endocrine-disrupting properties (Zhang et al., 2016). Another toxic chemical in our study, perfluorooctanesulfonate (PFOS), was 29% higher in relative abundance in urban versus rural participants and was the fourth important metabolite to distinguish urbanization status in random forest classification. PFOS has been shown to be adversely associated with body weight, thyroid function, and lipid and glucose metabolism in humans (Saikat et al., 2013) (Coperchini et al., 2017) and animal studies (Lai et al., 2018). Our findings show positive association between energy intake and diet diversity with 4-hydroxychlorothalonil and PFOS, indicating that water and food contaminations are the likely sources of these chemicals. Indeed, both chlorothalonil and PFOS were detected in water in China (Fang et al., 2019, Jin et al., 2009), with higher PFOS concentration in surface water in urban than rural areas (Chen et al., 2009) and chlorothalonil persisted in common produce like cabbage (Hou et al., 2016). In contrast, we found that physical activity was negatively associated with 4-hydroxychlorothalonil and PFOS. Our observed associations may relate to sweating during physical activity, as induced perspiration could accelerate the elimination of toxic chemicals like polychlorinated biphenyls in the human body (Genius et al., 2013).

The fact that xanthine metabolism pathways differed by urbanization status in our sample may reflect different urban and rural dietary habits, given that all of the detected metabolites from xanthine metabolism pathways are derived from caffeine and were positively associated with urbanization status. Caffeine is present in various foods and drinks, including coffee and tea that are more frequently consumed in urban versus rural residents in China (Zhang et al., 2011). One caffeine metabolite, theophylline, which is a bronchodilator that can treat asthma and other lung diseases, was higher in urban than rural participants, indicating that urban participants in our sample may be more likely to take this medication compared to rural participants. A potential reason for this difference is the heavier air pollution in urban versus rural areas, as a study in Switzerland has shown that traffic-related air pollution was positively associated with incidence of adult-onset asthma among never-smokers aged 18–60 years (Künzli et al., 2009).

In addition, we found that urbanization status was positively associated with metabolites involved in inflammation and metabolic disturbances, like leucine and isoleucine, two

BCAAs which have been shown to be related to increased risk of type 2 diabetes (Lotta et al., 2016, Newgard et al., 2009, Wang et al., 2011). Additionally, we found that the urbanization status was positively associated with cysteine-glutathione disulfide, a metabolite formed upon oxidative stress of glutathione. Oxidative stress has been shown to be involved in the development of cardiovascular diseases (CVD) (Alfadda and Sallam, 2012, Cervantes Gracia et al., 2017, Khosravi et al., 2019). Together, these results are consistent with the higher CVD risks in urban versus rural Chinese adults observed in a study of adults aged 35–70 years across 12 provinces between 2005 and 2009 (Yan et al., 2017) and a study of adults aged 45 years and older from 1991 to 2011 (Zhang, 2019). Conversely, we found that urbanization status was positively associated with a few metabolites potentially beneficial to cardiovascular health, such as alpha-tocopherol and gamma-tocopherol/beta-tocopherol, which are antioxidants with vitamin E activity that protect fatty acids from oxidation (Shahidi and De Camargo, 2016), as well as 1,2,3-benzenetriol sulfate, a gut microbiota-derived phenolic compound with anti-inflammatory properties (Larrosa et al., 2009), while negatively associated with Fibrinopeptide A, a well-established marker for coronary thrombosis (Eisenberg et al., 1985).

Our study has several strengths. First, we used the large, population-based CHNS, which provided metabolomics data from residents of two adjacent provinces, reducing confounding from geography and culture. Second, our data included detailed measures of environmental factors and behaviors, such as income, diet, and physical activity, allowing us to investigate the potential contributions of these factors to urban and rural differences in plasma metabolites. Third, we used the urbanization index that encompasses 12 diverse components of urbanization to categorize urban and rural areas, reducing potential misclassification by simple dichotomous characterization of urban and rural areas that uses only population density. Yet our study also has some limitations, including the cross-sectional design and relied on global metabolomics analysis only. For future identification of metabolite features assigned to Level 3 or Level 4 identification that contributed to the urban-rural separation in OPLS-DA (VIP 1.5) or random forest classification ( $G^2$  0.5), or nominally differed across urban and rural areas ( $p$ -value $<0.05$  in GLM), we provided the mass and retention index for these metabolites in Supplementary Table 8. It is possible that the accidental enrichment of metabolites with exogenous sources, such as tocopherols added into dietary supplements, may lead to misinterpretation of the pathway analysis. However, vitamin or mineral supplement use was uncommon in our participants; few urban ( $n=7$ ) and rural participants ( $n=1$ ) used supplements. We found that participants who took supplements within 24 h before blood sample collection, compared to those who did not, were not different in any of the detected cofactors and vitamins by urban (Supplementary Table 9,  $0.32$   $q$ -value  $0.81$ ) and rural areas ( $0.45$   $q$ -value  $0.81$ ), suggesting that it was unlikely that the enrichment of tocopherol metabolic pathway in urban versus rural participants was due to the use of supplements.

In conclusion, our findings suggest that urbanization status was associated with profound perturbation of host plasma metabolites in Chinese adults, including higher levels of hazardous chemicals, caffeine metabolites, and a few inflammatory metabolites in urban versus rural adults. These metabolites differences reflect changes in environment, lifestyle,

and diet related to urbanization. Our findings suggest that urbanization should be considered in metabolomic analysis.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGEMENTS

We are grateful to the participants in the China Health and Nutrition Survey. We also want to thank Dr. Hsiao-Chuan Tien and Ms. Guifeng Jin for database assistance.

### Funding

We are grateful to research grant funding from the National Institute for Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) for R01DK104371. We are also grateful to the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) for R01 HD30880 and the NIH Fogarty grant D43 TW009077 for financial support for the CHNS data collection and analysis files from 1989 to 2015 and future surveys. We thank the National Institute for Nutrition and Health, China Center for Disease Control and Prevention. We are also grateful for funding from the NICHD to Carolina Population Center at the University of North Carolina at Chapel Hill (NIH grant P2C HD050924). Y.W is grateful to the Sanofi funding award 29230-50347-466001 for educational support. M.C.B.T is supported by Ruth L. Kirschstein National Research Service Awards (NRSA) Genetic Epidemiology of Heart, Lung, and Blood (HLB) Traits Training Grant (GenHLB, T32HL129982) funded by National Heart, Lung, and Blood Institute (NHLBI).

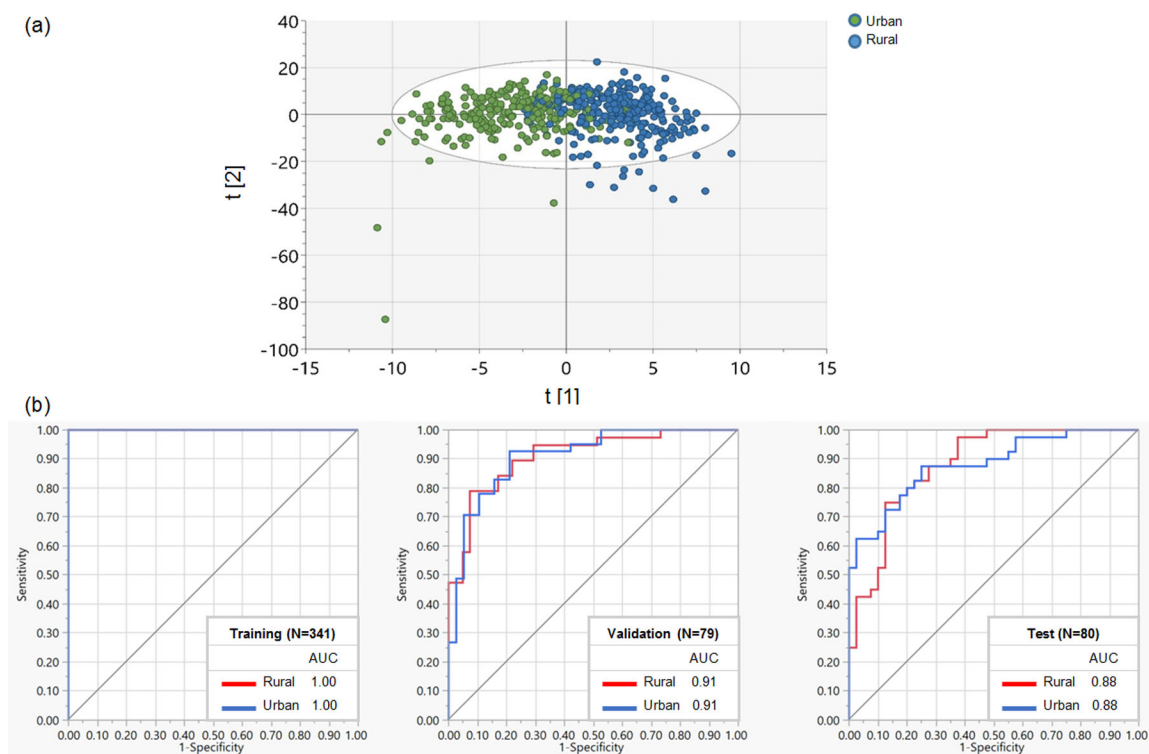
## REFERENCES

- ALFADDA AA & SALLAM RM 2012 Reactive oxygen species in health and disease. *Journal of Biomedicine and Biotechnology*, 2012.
- CERVANTES GRACIA K, LLANAS-CORNEJO D & HUSI H 2017 CVD and oxidative stress. *Journal of clinical medicine*, 6, 22.
- CHEN C, LU Y, ZHANG X, GENG J, WANG T, SHI Y, HU W & LI J 2009 A review of spatial and temporal assessment of PFOS and PFOA contamination in China. *Chemistry and Ecology*, 25, 163–177.
- COPERCHINI F, AWWAD O, ROTONDI M, SANTINI F, IMBRIANI M & CHIOVATO L 2017 Thyroid disruption by perfluorooctane sulfonate (PFOS) and perfluorooctanoate (PFOA). *Journal of endocrinological investigation*, 40, 105–121. [PubMed: 27837466]
- DING M, ZELEZNIK OA, GUASCH-FERRE M, HU J, LASKY-SU J, LEE I-M, JACKSON RD, SHADYAB AH, LAMONTE MJ & CLISH C 2019 Metabolome-wide Association Study with Habitual Physical Activity in Four Prospective Cohort Studies. *American journal of epidemiology*.
- EISENBERG PR, SHERMAN LA, SCHECTMAN K, PEREZ J, SOBEL B & JAFFE AS 1985 Fibrinopeptide A: a marker of acute coronary thrombosis. *Circulation*, 71, 912–918. [PubMed: 3986981]
- EVANS AM, BRIDGEWATER B, LIU Q, MITCHELL M, ROBINSON R, DAI H, STEWART S, DEHAVEN C & MILLER L 2014 High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. *Metabolomics*, 4(2), 1.
- EVANS AM, DEHAVEN CD, BARRETT T, MITCHELL M & MILGRAM E 2009 Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical chemistry*, 81, 6656–6667. [PubMed: 19624122]
- FANG W, PENG Y, MUIR D, LIN J & ZHANG X 2019 A critical review of synthetic chemicals in surface waters of the US, the EU and China. *Environment international*, 131, 104994. [PubMed: 31302480]

- GENUIS SJ, BEESOON S & BIRKHOLZ D 2013 Biomonitoring and elimination of perfluorinated compounds and polychlorinated biphenyls through perspiration: blood, urine, and sweat study. *ISRN toxicology*, 2013.
- HILTUNEN TP, RIMPELÄ JM, MOHNEY RP, STIRDIVANT SM & KONTULA KK 2017 Effects of four different antihypertensive drugs on plasma metabolomic profiles in patients with essential hypertension. *PLoS one*, 12, e0187729. [PubMed: 29121091]
- HOU F, ZHAO L & LIU F 2016 Residues and dissipation of chlorothalonil and azoxystrobin in cabbage under field conditions. *International journal of environmental analytical chemistry*, 96, 1105–1116.
- INOUE Y, HOWARD AG, THOMPSON AL & GORDON-LARSEN P 2018 Secular change in the association between urbanisation and abdominal adiposity in China (1993–2011). *J Epidemiol Community Health*, 72, 484–490. [PubMed: 29514924]
- JIN YH, LIU W, SATO I, NAKAYAMA SF, SASAKI K, SAITO N & TSUDA S 2009 PFOS and PFOA in environmental and tap water in China. *Chemosphere*, 77, 605–611. [PubMed: 19775722]
- JONES-SMITH JC & POPKIN BM 2010 Understanding community context and adult health changes in China: development of an urbanicity scale. *Social science & medicine*, 71, 1436–1446. [PubMed: 20810197]
- KHOSRAVI M, POURSALEH A, GHASEMPOUR G, FARHAD S & NAJAFI M 2019 The effects of oxidative stress on the development of atherosclerosis. *Biological chemistry*, 400, 711–732. [PubMed: 30864421]
- KÜNZLI N, BRIDEVAUX P-O, LIU LS, GARCIA-ESTEBAN R, SCHINDLER C, GERBASE M, SUNYER J, KEIDEL D & ROCHAT T 2009 Traffic-related air pollution correlates with adult-onset asthma among never-smokers. *Thorax*, 64, 664–670. [PubMed: 19359271]
- LAI KP, NG AH-M, WAN HT, WONG AYM, LEUNG CCT, LI R & WONG CK 2018 Dietary exposure to the environmental chemical, PFOS on the diversity of gut microbiota, associated with the development of metabolic syndrome. *Frontiers in microbiology*, 9, 2552. [PubMed: 30405595]
- LARROSA M, LUCERI C, VIVOLI E, PAGLIUCA C, LODOVICI M, MONETI G & DOLARA P 2009 Polyphenol metabolites from colonic microbiota exert anti-inflammatory activity on different inflammation models. *Mol Nutr Food Res*, 53, 1044–54. [PubMed: 19557820]
- LIN S, SUN J, MARINOVA D & ZHAO D 2017 Effects of population and land urbanization on China's environmental impact: Empirical analysis based on the extended STIRPAT model. *Sustainability*, 9, 825.
- LOTTA LA, SCOTT RA, SHARP SJ, BURGESS S, LUAN JA, TILLIN T, SCHMIDT AF, IMAMURA F, STEWART ID & PERRY JR 2016 Genetic predisposition to an impaired metabolism of the branched-chain amino acids and risk of type 2 diabetes: a Mendelian randomisation analysis. *PLoS medicine*, 13, e1002179. [PubMed: 27898682]
- MAKRECKA M, KUKA J, VOLSKA K, ANTONE U, SEVOSTJANOV S, CIRULE H, GRINBERGA S, PUGOVICS O, DAMBROVA M & LIEPINSH E 2014 Long-chain acylcarnitine content determines the pattern of energy metabolism in cardiac mitochondria. *Molecular and cellular biochemistry*, 395, 1–10. [PubMed: 24878991]
- MCMAHON TA, HALSTEAD NT, JOHNSON S, RAFFEL TR, ROMANSIC JM, CRUMRINE PW, BOUGHTON RK, MARTIN LB & ROHR JR 2011 The fungicide chlorothalonil is nonlinearly associated with corticosterone levels, immunity, and mortality in amphibians. *Environmental health perspectives*, 119, 1098–1103. [PubMed: 21463979]
- MEHMOOD T, LILAND KH, SNIPEN L & SÆBØ S 2012 A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118, 62–69.
- MENNI C, GRAHAM D, KASTENMÜLLER G, ALHARBI NH, ALSANOSI SM, MCBRIDE M, MANGINO M, TITCOMBE P, SHIN S-Y & PSATHA M 2015 Metabolomic identification of a novel pathway of blood pressure regulation involving hexadecanedioate. *Hypertension*, 66, 422–429. [PubMed: 26034203]
- NEWGARD CB, AN J, BAIN JR, MUEHLBAUER MJ, STEVENS RD, LIEN LF, HAQQ AM, SHAH SH, ARLOTTO M & SLENTZ CA 2009 A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell metabolism*, 9, 311–326. [PubMed: 19356713]

- NG SW, HOWARD AG, WANG H, SU C & ZHANG B 2014 The physical activity transition among adults in China: 1991–2011. *Obesity Reviews*, 15, 27–36.
- NG SW & POPKIN BM 2012 Time use and physical activity: a shift away from movement across the globe. *Obesity reviews*, 13, 659–680. [PubMed: 22694051]
- POPKIN BM, ADAIR LS & NG SW 2012 Global nutrition transition and the pandemic of obesity in developing countries. *Nutrition reviews*, 70, 3–21. [PubMed: 22221213]
- POPKIN BM & DU S 2003 Dynamics of the nutrition transition toward the animal foods sector in China and its implications: a worried perspective. *The Journal of nutrition*, 133, 3898S–3906S. [PubMed: 14672288]
- POPKIN BM, DU S, ZHAI F & ZHANG B 2009 Cohort Profile: The China Health and Nutrition Survey—monitoring and understanding socio-economic and health change in China, 1989–2011. *International journal of epidemiology*, 39, 1435–1440. [PubMed: 19887509]
- ROSSI LA 1999 Reregistration eligibility decision (RED) chlorothalonil. UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (EPA), Special Review and Reregistration Division (7508C).
- SAIKAT S, KREIS I, DAVIES B, BRIDGMAN S & KAMANYIRE R 2013 The impact of PFOS on health in the general population: a review. *Environmental Science: Processes & Impacts*, 15, 329–335. [PubMed: 25208696]
- SHAHIDI F & DE CAMARGO AC 2016 Tocopherols and tocotrienols in common and emerging dietary sources: Occurrence, applications, and health benefits. *International journal of molecular sciences*, 17, 1745.
- SHIN J-H, JUNG S, KIM S-A, KANG M-S, KIM M-S, JOUNG H, HWANG G-S & SHIN D-M 2019 Differential Effects of Typical Korean Versus American-Style Diets on Gut Microbial Composition and Metabolic Profile in Healthy Overweight Koreans: A Randomized Crossover Trial. *Nutrients*, 11, 2450.
- STOREY JD 2002 A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 479–498.
- STOREY JD, TAYLOR JE & SIEGMUND D 2004 Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 187–205.
- SUMNER LW, AMBERG A, BARRETT D, BEALE MH, BEGER R, DAYKIN CA, FAN TW-M, FIEHN O, GOODACRE R & GRIFFIN JL 2007 Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3, 211–221. [PubMed: 24039616]
- VAN VELDHOVEN K, KISS A, KESKI-RAHKONEN P, ROBINOT N, SCALBERT A, CULLINAN P, CHUNG KF, COLLINS P, SINHARAY R & BARRATT BM 2019 Impact of short-term traffic-related air pollution on the metabolome—Results from two metabolome-wide experimental studies. *Environment international*, 123, 124–131. [PubMed: 30522001]
- WANG TJ, LARSON MG, VASAN RS, CHENG S, RHEE EP, MCCABE E, LEWIS GD, FOX CS, JACQUES PF & FERNANDEZ C 2011 Metabolite profiles and the risk of developing diabetes. *Nature medicine*, 17, 448.
- YAN R, LI W, YIN L, WANG Y, BO J, INVESTIGATORS PC, INVESTIGATORS PC, LIU L, LIU B & HU B 2017 Cardiovascular diseases and risk-factor burden in urban and rural communities in high-, middle-, and low-income regions of China: a large community-based epidemiological study. *Journal of the American Heart Association*, 6, e004445. [PubMed: 28167497]
- YANG Y-X 2005 Chinese food composition table 2004. Beijing: Peking University Medical Press.
- YAO M, LICHTENSTEIN A, ROBERTS S, MA G, GAO S, TUCKER K & MCCRORY M 2003 Relative influence of diet and physical activity on cardiovascular risk factors in urban Chinese adults. *International journal of obesity*, 27, 920. [PubMed: 12861233]
- YONESHIRO T, WANG Q, TAJIMA K, MATSUSHITA M, MAKI H, IGARASHI K, DAI Z, WHITE PJ, MCGARRAH RW & ILKAYEVA OR 2019 BCAA catabolism in brown fat controls energy homeostasis through SLC25A44. *Nature*, 572, 614–619. [PubMed: 31435015]
- ZHANG B, WEN-WEN D, WANG H, CHANG S, ZHANG J & ZHAI F 2011 Tea and coffee consumption status and trend among 18 to 49 years old adults in 9 provinces of China. *Food & Nutrition in China*, 4.

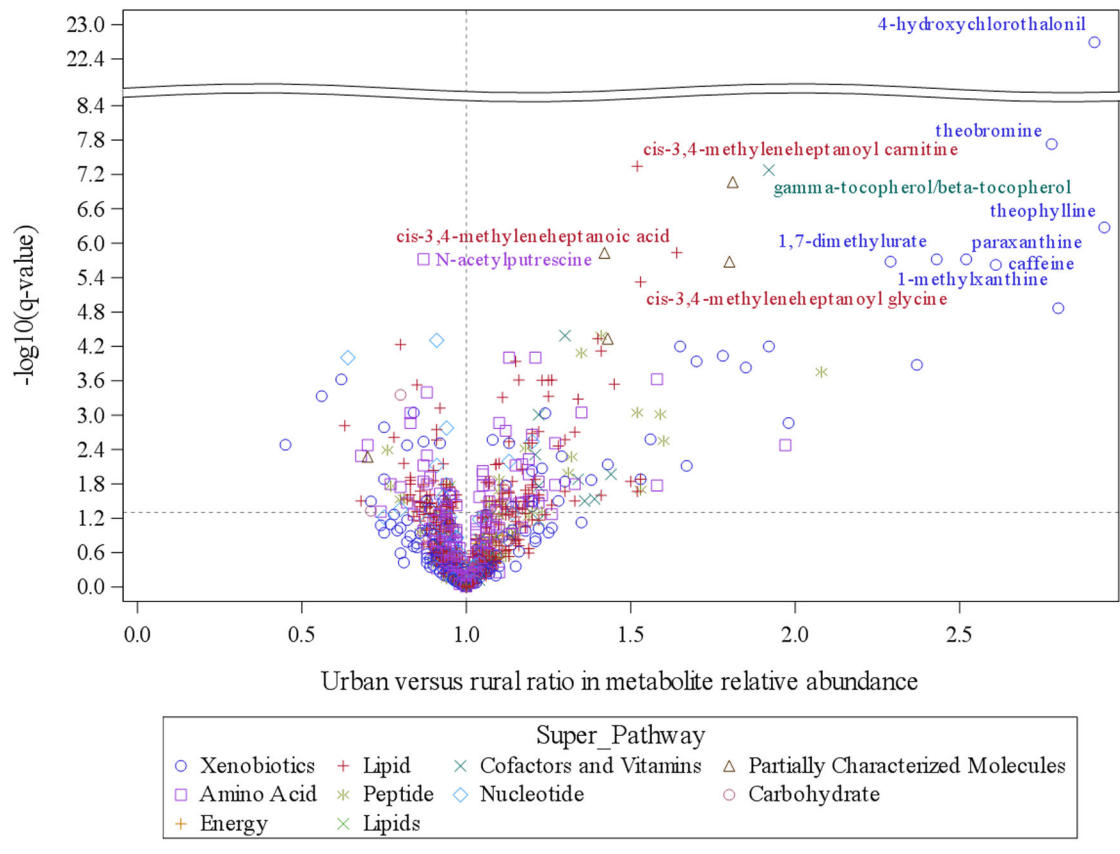
- ZHANG N 2019 Urban–rural disparities in cardiovascular disease risks among middle-aged and older Chinese: two decades of urbanisation. *Ageing & Society*, 1–23. [PubMed: 32863475]
- ZHANG Q, JI C, YAN L, LU M, LU C & ZHAO M 2016 The identification of the metabolites of chlorothalonil in zebrafish (*Danio rerio*) and their embryo toxicity and endocrine effects at environmentally relevant levels. *Environmental pollution*, 218, 8–15. [PubMed: 27541960]
- ZHERNAKOVA DV, LE TH, KURILSHIKOV A, ATANASOVSKA B, BONDER MJ, SANNA S, CLARINGBOULD A, VOSA U, DEELEN P & FRANKE L 2018 Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nature genetics*, 50, 1524–1532. [PubMed: 30250126]
- ZHOU M, JING S, WU C-Y, SHU L, DONG W, LIU Y, CHEN M, WYNN RM, WANG J & WANG J 2019 Targeting BCAA Catabolism to Treat Obesity-Associated Insulin Resistance. *Diabetes*, db180927.

**Fig 1.**

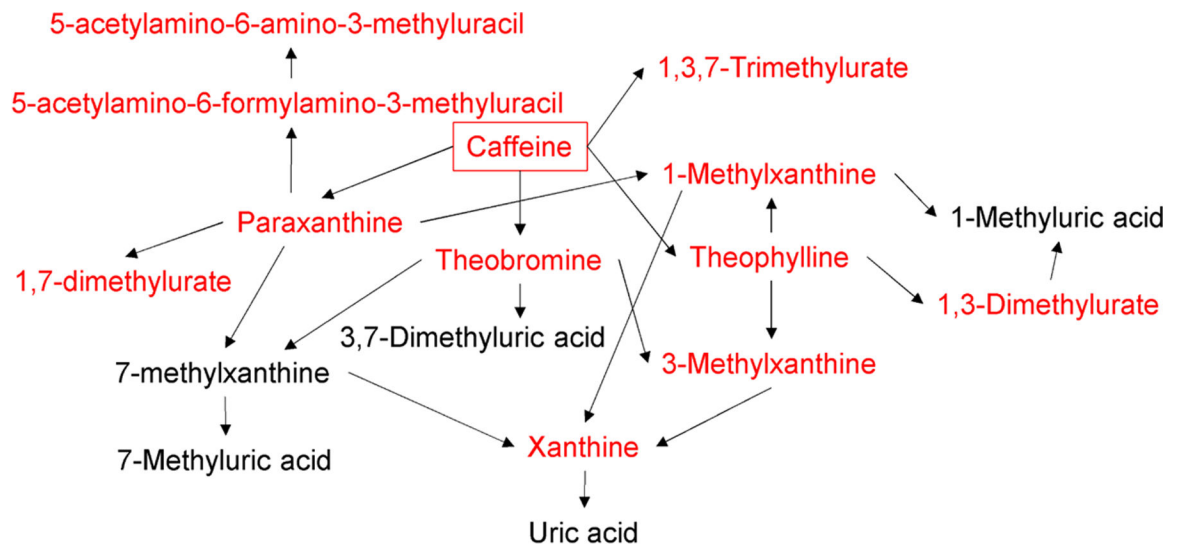
**(a) Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) score plot for the separation of plasma metabolome by urban and rural status.** T [1] and t [2] are the first two latent variables (i.e., linear combinations of metabolite data) in the OPLS-DA model using all identified metabolite features, where t [1] is the class-predictive latent variable that captures the variation in metabolites between urban and rural areas and t [2] is the class-orthogonal latent variable that includes systematic variation in metabolite data uncorrelated with urbanization status. Together, t [1] and t [2] provide the best separation of the urban and rural samples in OPLS-DA. **(b) Receiver operating characteristic (ROC) curve for the discrimination of urban and rural metabolome using random forest classification.**

AUC, area under the curve.



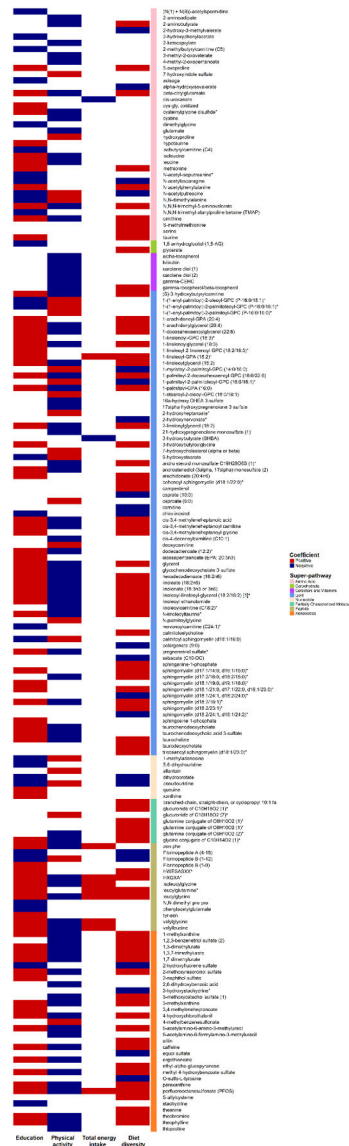


**Fig 2. Volcano plot for the association between urbanization and individual metabolites.** X axis is the urban versus rural ratio of metabolites relative abundance and y axis is the  $-\log_{10}(\text{q-value})$  based on linear regression adjusting for age, sex, and province. Metabolites types (super-pathway) are indicated by different colors and shapes. Metabolites with  $\text{q-value} < 1.0\text{E-}05$  are labeled (excluding three partially characterized metabolites).



**Fig 3. Biochemical pathway of xanthine metabolism.**

Urbanization status was positively associated with all of the detected metabolites in xanthine metabolism ( $q$ -value $<0.05$ ), which are indicated in red. Metabolites colored in black were not detected in this study. Note that xanthine was categorized to purine metabolism per Metabolon's definition.



**Fig 4. The association between key urbanization-related factors and urbanization-associated with metabolites.**

Direction of associations and statistical significance are indicated by color and shading, respectively. Metabolites are ordered by types (super-pathways), which are differentiated by color. Models were adjusted for age, sex, and province. A total of 266 metabolites were tested. Metabolites that were not associated with any of the factors (q-value 0.05) are not listed in the figure. Per capita household income and animal-source food consumption were not associated with any of the metabolites at q-value<0.05 and thus not shown in the figure. \* Metabolites at Level 2 identification according to the Metabolomics Standards Initiative.

**Table 1.**

Characteristics of the China Health and Nutrition Survey participants with plasma metabolomic data, median (interquartile range) or %

	Urban <sup>a</sup>	Rural <sup>a</sup>	P-value <sup>b</sup>
N	240	260	
Age	54.6 (46.4–60.3)	51.8 (44.3–60.1)	0.193
Women	57.9	60.4	0.575
Completed high school education	46.7	15.8	<0.0001
Per capita household income (¥1000)	20.3 (12.4–34.0)	12.8 (5.1–21.5)	<0.0001
Physical activity (METS/wk) <sup>c</sup>	63.8 (18.9–111.8)	130.5 (57.8–253.3)	<0.0001
Total energy intake (kcal) <sup>d</sup>	1872.8 (1435.3–2293.7)	1744.3 (1400.7–2188.79)	0.112
Animal-source foods (%kcal) <sup>d</sup>	24.0 (16.7–32.7)	24.0 (14.8–33.9)	0.862
Diet diversity <sup>e</sup>	12.0 (10.0–14.0)	10.0 (8.0–12.0)	<0.0001

<sup>a</sup>We used a 12-component index with a continuous scale of 0–120 points to assess multiple aspects of urbanization, including population density, economic activity, transportation infrastructure, and sanitation. We dichotomized this urbanization index by median (=73.1) to define urban (76.1–99.6) and rural areas (39.2–73.1).

<sup>b</sup>Urban-rural differences in categorical variables were assessed with  $\chi^2$  and continuous variables were assessed with Wilcoxon rank sum test.

<sup>c</sup>Physical activity in metabolic equivalent of tasks (METs) per week was measured by 7-day recalls of all domestic, occupational, transportation, and leisure activities.

<sup>d</sup>Nutrients intake was estimated using 3 consecutive 24-h recalls and household food inventories by trained interviewers, with total energy intake validated by doubly labeled water.

<sup>e</sup>A count of number of food groups consumed, including staple grains, fruits, vegetables, nuts, eggs, meat, seafood, and dairy products.

**Table 2.**

## Pathway enrichment analysis

Sub-pathway	k <sup>a</sup>	m <sup>a</sup>	Enrichment Score <sup>a</sup>	P-value <sup>b</sup>	Q-value <sup>b</sup>
Xanthine Metabolism	11	11	3.45	1.47E-06	1.64E-04
Sphingomyelins	16	27	2.05	0.001	0.068
Partially Characterized Molecules	7	9	2.65	0.004	0.150
Glutathione Metabolism	5	7	2.42	0.028	0.480
Dipeptide	8	14	1.94	0.030	0.480
Fatty Acid Metabolism (Acyl Glycine)	4	5	2.70	0.030	0.480
Tocopherol Metabolism	4	5	2.70	0.030	0.480
Pregnenolone Steroids	5	8	2.11	0.056	0.788
Carnitine Metabolism	2	2	3.36	0.089	0.998
Polypeptide	2	2	3.36	0.089	0.998
Polyamine Metabolism	4	7	1.93	0.124	1.000
Phosphatidylcholine (PC)	8	18	1.50	0.136	1.000
Glycine, Serine and Threonine Metabolism	5	10	1.69	0.147	1.000
Long Chain Polyunsaturated Fatty Acid (n3 and n6)	7	16	1.48	0.171	1.000
Chemical	9	22	1.38	0.181	1.000
Tobacco Metabolite	2	3	2.24	0.214	1.000
Leucine, Isoleucine and Valine Metabolism	11	30	1.24	0.262	1.000
Fatty Acid, Amide	1	1	3.35	0.299	1.000
Ketone Bodies	1	1	3.35	0.299	1.000
Modified Peptides	1	1	3.35	0.299	1.000
Hemoglobin and Porphyrin Metabolism	2	4	1.68	0.346	1.000
Vitamin A Metabolism	2	4	1.68	0.346	1.000
Sterol	3	7	1.44	0.350	1.000
Gamma-glutamyl Amino Acid	6	17	1.19	0.399	1.000
Monoacylglycerol	5	14	1.20	0.412	1.000
Medium Chain Fatty Acid	4	11	1.22	0.427	1.000
Plasmalogen	4	11	1.22	0.427	1.000
Fatty Acid Metabolism (Acyl Choline)	3	8	1.26	0.446	1.000
Food Component/Plant	17	54	1.06	0.449	1.000
Dihydro sphingomyelins	2	5	1.34	0.470	1.000
Fatty Acid Metabolism (Acyl Carnitine, Polyunsaturated)	2	5	1.34	0.470	1.000
Glycolysis, Gluconeogenesis, and Pyru..	2	5	1.34	0.470	1.000
Purine Metabolism, Adenine containing	2	5	1.34	0.470	1.000
Methionine, Cysteine, SAM and Taurine..	7	22	1.07	0.502	1.000
Inositol Metabolism	1	2	1.68	0.509	1.000
Sphingolipid Synthesis	1	2	1.68	0.509	1.000
Sphingosines	1	2	1.68	0.509	1.000
Fatty Acid Metabolism (Acyl Carnitine, Monounsaturated)	3	9	1.12	0.535	1.000
Fibrinogen Cleavage Peptide	3	9	1.12	0.535	1.000

Sub-pathway	k <sup>a</sup>	m <sup>a</sup>	Enrichment Score <sup>a</sup>	P-value <sup>b</sup>	Q-value <sup>b</sup>
Fatty Acid Metabolism (Acyl Carnitine, Medium Chain)	2	6	1.12	0.578	1.000
Tryptophan Metabolism	6	20	1.00	0.581	1.000
Fatty Acid, Monohydroxy	5	17	0.98	0.609	1.000
Primary Bile Acid Metabolism	3	10	1.00	0.615	1.000
Urea cycle; Arginine and Proline Metabolism	6	21	0.95	0.635	1.000
Acetylated Peptides	1	3	1.12	0.656	1.000
Diacylglycerol	1	3	1.12	0.656	1.000
Fatty Acid, Amino	1	3	1.12	0.656	1.000
Fatty Acid, Branched	1	3	1.12	0.656	1.000
Glycerolipid Metabolism	1	3	1.12	0.656	1.000
Pyrimidine Metabolism, Orotate containing	1	3	1.12	0.656	1.000
Endocannabinoid	2	7	0.96	0.669	1.000
Purine Metabolism, (Hypo)Xanthine/Inosine containing	2	7	0.96	0.669	1.000
Glutamate Metabolism	3	11	0.91	0.686	1.000
Pyrimidine Metabolism, Uracil containing	3	11	0.91	0.686	1.000
Benzoate Metabolism	8	30	0.89	0.718	1.000
Alanine and Aspartate Metabolism	2	8	0.84	0.744	1.000
Phenylalanine Metabolism	2	8	0.84	0.744	1.000
Histidine Metabolism	4	16	0.83	0.753	1.000
Eicosanoid	1	4	0.84	0.759	1.000
Fatty Acid Metabolism (Acyl Carnitine, Hydroxy)	1	4	0.84	0.759	1.000
Fructose, Mannose and Galactose Metabolism	1	4	0.84	0.759	1.000
Purine Metabolism, Guanine containing	1	4	0.84	0.759	1.000
Phosphatidylinositol (PI)	1	5	0.67	0.831	1.000
Fatty Acid, Dicarboxylate	7	31	0.75	0.867	1.000
Fatty Acid Metabolism (Acyl Carnitine, Long Chain Saturated)	1	6	0.56	0.882	1.000
Androgenic Steroids	4	21	0.63	0.915	1.000
Long Chain Monounsaturated Fatty Acid	1	7	0.48	0.918	1.000
Nicotinate and Nicotinamide Metabolism	1	7	0.48	0.918	1.000
Lysophospholipid	4	22	0.60	0.933	1.000
Phosphatidylethanolamine (PE)	1	8	0.42	0.942	1.000
Lysine Metabolism	2	15	0.44	0.965	1.000
Tyrosine Metabolism	1	17	0.19	0.998	1.000
Secondary Bile Acid Metabolism	2	24	0.27	0.998	1.000

<sup>a</sup> Enrichment score was calculated using  $(k/m)/[(n-k)/(N-m)]$ , where k is the number of significant metabolites [positive false discovery rate (pFDR) adjusted p-value, q-value<0.05] in the sub-pathway, m is the total number of detected metabolites in the pathway, n is the total number of significant (q-value<0.05) and classified metabolites (n=266), and N is the total number of detected and classified metabolites (N=890). A total of 112 metabolic pathways were tested. Only metabolic pathways with k>0 are shown here.

<sup>b</sup> P-value for each sub-pathway was calculated using Fisher's exact test and pFDR adjusted.