Editor Recognition: Primary Handling editor: Yanina-Yasmin Pesch, in collaboration with the Nature Metabolism team.

Peer review Information:
Nature Metabolism thanks Rui Wang-Sattler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Editor Summary Note 'In a large multi-ethnic Asian cohort, associations between over 1000 plasma metabolites and specific foods and beverages are made. These diet-metabolite relationships were used to accurately predict clinical phenotypes such as diabetes and hypertension

Figure or Table # Please group Extended Data items by type, in sequential order. Total number of items (Figs. + Tables) must not exceed 10.	Figure/Table title One sentence only	Filename Whole original file name including extension. i.e.: Smith_ED_Fig1.jpg	Figure/Table Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Table. 1	Extended Data Table 1	EDT 1.pdf	Characteristics of HELIOS study population at baseline visit. Values are presented as mean±SD.
Extended Data Table. 2	Extended Data Table 2	EDT 2.pdf	Characteristics of subset of HELIOS study population at baseline and repeat visits. Values are presented as mean±SD.
Extended Data Table. 3	Extended Data Table 3	EDT 3.pdf	Description of 42 food and beverage groups, where multi-biomarker panels were proposed for 20 key foods/beverages[1] for in the HELIOS population study.
Extended Data Fig. 1	Extended Data Fig 1	ExtendedDataFig 1.jpg	Indices of diet quality components within the HELIOS study population. a) Sankey illustration of diet quality components. b) Radar plots showing distribution of individual components across diet quality indices. c) Density plots of diet quality scores across age and ethnicity. MUFA, monounsaturated fatty acids; PUFA, polyunsaturated fatty acids; SFA, saturated fatty

			acids; bev, beverage; proc, processed; HF, high fat; LF, low fat.
Extended Data Fig. 2	Extended Data Fig 2	ExtendedDataFig 2.jpg	Model prediction performance of dietary intakes and diet quality. a) Percent (%) error observed between self-reported and estimated intakes (log-transformed) of foods and beverages in respective test datasets (n=2,336-2,513; refer to Supplementary Table 7 for exact sample sizes of test datasets). b) Density distribution of prediction error % of foods and beverages in respective test datasets with dotted line indicating median error and text insert indicating median prediction accuracy (formula: 1 - error %).
Extended Data Fig. 3	Extended Data Fig 3	ExtendedDataFig 3.jpg	Summary of coffee and tea intakes. a) Percent (%) of variations of total daily coffee and tea intakes recorded via FFQ in the study population. b) Proportion of kilocalories across the top 5 most consumed variations of coffee or tea.
Extended Data Fig. 4	Extended Data Fig 4	ExtendedDataFig 4.jpg	Associations between environmental contaminants and clinical phenotypes. Effect sizes of environmental contaminants (that comprise various dietary biomarker panels), adjusted for age, sex, ethnicity and batch. Estimates were obtained from independent linear regression (for HOMA-IR, FMI, cIMT) or logistic regression (for T2D, hypertension), based on the following sample sizes: T2D (n=1,040 T2D, n=7,351 non-T2D), hypertension (n=2,411 with hypertension, n=5,980 without hypertension). Association strength represents log-transformed *P* values. Circle border represents significant associations ($P<0.05$). HOMA-IR, homeostatic model assessment for insulin resistance; T2D, type 2

			diabetes; FMI, fat mass index; cIMT, carotid intima-media thickness.

12

Item	Present?	Filename Whole original file name including extension. i.e.: Smith_SI.pdf. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary_Tables.xlsx	Supplementary Tables 1-10
Reporting Summary	Yes	Chambers_RS flat.pdf	
Peer Review Information	No	*OFFICE USE ONLY*	

13
14
15
16

Type	Number Each type of file (Table, Video, etc.) should be numbered from 1 onwards. Multiple files of the same type should be listed in sequence, i.e.: Supplementary Video 1, Supplementary Video 2, etc.	Filename Whole original file name including extension. i.e.: *Smith_ Supplementary_Video_1.mov*	Legend or Descriptive Caption Describe the contents of the file
Supplementary Table	**Supplementary Tables 1-10**	Supplementary_Tables.xlsx	Supplementary Table 1: List of 1,055 metabolites in the study population. Supplementary Table 2: Partial correlation coefficients of FFQ food/beverage items and metabolites ($r \geq 0.15$ and

			FDR-*P*<0.05), adjusted for age, sex, ethnicity and batch. Supplementary Table 3: Effect sizes of metabolites (within the multi-biomarker panels) from associations with key foods and beverages, in regression models adjusted for age, sex, batch and ethnicity or genetic ancestry. Supplementary Table 4: Summary table of metabolite and genetic correlations for annotation of unknowns in the dietary biomarker panels. Supplementary Table 5: Summary of metabolic pathways enriched in dietary quality indices. Supplementary Table 6: Effect sizes of nutrients (% kcal) from associations with dietary quality indices. Supplementary Table 7: Effect sizes of dietary

			intakes (derived from self-report or metabolite score) from associations with clinical phenotypes. Supplementary Table 8: Sample sizes for training and test datasets of each food and beverage dataset. Supplementary Table 9: Effect sizes of diet quality (derived from self-report or metabolite score) from associations with clinical phenotypes.

17

18

Parent Figure or Table	Filename Whole original file name including extension. i.e.: *Smith_SourceData_Fig1.xls,* or *Smith_ Unmodified_Gels_Fig1.pdf*	Data description i.e.: Unprocessed western Blots and/or gels, Statistical Source Data, etc.
Source Data Fig. 1	SourceData_Fig1.xlsx	Statistical source data
Source Data Fig. 2	SourceData_Fig2.xlsx	Statistical source data
Source Data Fig. 3	SourceData_Fig3.xlsx	Statistical source data
Source Data Fig. 4	SourceData_Fig4.xlsx	Statistical source data
Source Data Fig. 5	SourceData_Fig5.xlsx	Statistical source data
Source Data Fig. 6	SourceData_Fig6.xlsx	Statistical source data
Source Data Extended Data Fig. 1	SourceData_ExtendedFig1.xlsx	Statistical source data

Source Data Extended Data Fig. 2	SourceData_ExtendedFig2.xlsx	Statistical source data
Source Data Extended Data Fig. 3	SourceData_ExtendedFig3.xlsx	Statistical source data
Source Data Extended Data Fig. 4	SourceData_ExtendedFig4.xlsx	Statistical source data
Source Data Extended Data Table 1	SourceData_ExtendedTable1.xlsx	Statistical source data
Source Data Extended Data Table 2	SourceData_ExtendedTable2.xlsx	Statistical source data
Source Data Extended Data Table 3	SourceData_ExtendedTable3.xlsx	Statistical source data

19

Metabolic variation reflects dietary exposure in a multi-ethnic Asian population

Dorrain Y Low[1]*, Theresia H Mina[1], Nilanjana Sadhu[1], Kari E Wong[2], Pritesh Rajesh Jain[1], Rinkoo Dalan[1,3], Hong Kiat Ng[1], Wubin Xie[1], Benjamin Lam[1,4], Darwin Tay[1], Xiaoyan Wang[1], Yik Weng Yew[1,5], James D Best[1], Rangaprasad Sarangarajan[2], Paul Elliott[1,6], Elio Riboli[1,6], Jimmy Lee[1,7], Eng Sing Lee[1,8], Joanne Ngeow[1,9], Patricia A Sheridan[2], Xue Li Guan[1], Gregory A Michelotti[2], Marie Loh[1,5], John C Chambers[1,6]*

[1]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
[2]Metabolon, Morrisville, North Carolina, USA
[3]Department of Endocrinology, Tan Tock Seng Hospital, Singapore
[4]Integrated Care for Obesity & Diabetes, Khoo Teck Puat Hospital, Singapore
[5]National Skin Centre, Research Division, Singapore
[6]Department of Epidemiology and Biostatistics, Imperial College London, London, UK
[7]North Region, Institute of Mental Health, Singapore
[8]Clinical Research Unit, National Healthcare Group Polyclinic, Singapore
[9]Division of Medical Oncology, National Cancer Centre, Singapore

*Corresponding authors:
John Chambers; Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore; john.chambers@ntu.edu.sg; +65 69041299
Dorrain Yanwen Low; Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore; dorrain.low@ntu.edu.sg

Abstract (150 words)

Understanding how diet shapes metabolism across diverse populations is essential to improving nutrition and health. Biomarkers reflecting diet are explored largely in European and American populations, but the food metabolome is highly complex, and varies across region and culture. We assessed 1,055 plasma metabolites and 169 foods/beverages in 8,391 multi-ethnic Asian individuals and carried out diet-metabolite association analyses. Using machine-learning, we developed multi-biomarker panels and composite scores for key foods, beverages and overall diet quality. Here we show these biomarker panels can be used to objectively assess dietary intakes in the Asian multi-ethnic population and can explain variances in intake prediction models better than single biomarkers. The identified diet-metabolite relationships are reproducible over time and improve prediction of clinical outcomes (insulin resistance, diabetes, BMI, carotid intima-media thickness and hypertension), comparing to self-reports. Our findings show insights into multi-ethnic diet-related metabolic variations and an opportunity to link exposure to population health outcomes.

Introduction

There is a growing global[1] pandemic of non-communicable diseases (NCDs) including obesity, diabetes and cardiovascular disease. Although there are established genetic and environmental contributors to NCDs mortality and morbidity, modifiable lifestyle determinants such as diet, play a major role[2,3]. The marked rise in NCDs has a causal link to global dietary habits becoming increasingly urbanised, being characterised by higher intakes of processed foods, refined grains and saturated fat, and concomitant lack of fibre-rich foods[4]. The Global Burden of Disease Study[5] highlighted suboptimal diets accounts for >10% of global deaths (more deaths than any other NCD risk factors including smoking). This emphasises the importance of accurate dietary assessment and their relationship to health.

Questionnaires are conventionally used to record individual-level dietary habits. The limitations of questionnaire based approaches are well-documented including recall bias, poor reproducibility, social desirability and restricted list of foods and/or beverages[6,7],which are compounded by variability in food composition databases. When questionnaire-based dietary data are linked to diet-disease relationships, this may cause estimate errors and attenuation of relationships. Even if dietary questionnaires have been pivotal in large population studies to guide development of nutrition recommendations or policy formulations, these shortcomings lead to concerns about the quality of evidence in nutritional epidemiology and evidence for dietary interventions.

Profiling our “food metabolome” using untargeted metabolomics approaches has emerged as a valuable strategy in discovery of dietary biomarkers[8] as objective assessment of habitual dietary exposures. This approach has created awareness of the untracked diversity of compounds in foods that remains largely invisible to epidemiological or hypothesis-driven nutrition studies[9]. These diet-derived metabolites reflect true concentrations of ingested foods and/or their metabolic products in biofluids (blood, urine), and represent an opportunity to overcome the limitations of questionnaires, while simultaneously accounting for interindividual variability in metabolism including genetic background and gut-microbial composition[10,11]. Dietary biomarkers (either as single metabolite or combination of metabolites) have been identified and validated for coffee[12], orange[13], banana[14] and red meat[15] in European and American populations but are lacking for other foods, foods consumed as part of complex dietary patterns and/or relevant to populations in other regions. For instance, different coffee biomarkers, rather than a unique biomarker, were identified across four European populations[16], corresponding to coffee preparation in each country. Considering regional and cultural diet diversity across the world, biomarkers relevant to Asian populations or overall diet quality are lacking.

In this study, we aim to i) identify relevant dietary biomarker panels reflecting food and beverage intakes, and ii) diet quality indices relevant a multi-ethnic Asian population, and iii) derive composite metabolite scores to improve predictions of diet-disease relationships.

Results

Study design

Baseline dietary, clinical and biological measurements were collected from 10,004 free-living adults (30-84 years old) from the multi-ethnic Singapore population in the HELIOS

study (**Fig. 1; Extended Data Table 1**). We explored plasma metabolomics profiles of habitual dietary exposures in an Asian epidemiological setting, and derived multi-biomarker panels using machine learning models, to demonstrate the potential utility of these panels as an objective and reproducible assessment of dietary intakes and quality. We also derived composite metabolite scores to improve predictions of diet and cardiometabolic disease relationships.

Plasma metabolome reflects self-reported dietary exposure

Plasma metabolomics profiling by mass spectrometry revealed 1,055 distinct metabolites (**Supplementary Table 1**), of which 887 (80%) are structurally identified. We found 156 metabolites positively correlated with 37 foods and beverages after adjusting for age, sex, ethnicity and batch (r=0.15-0.68, FDR P<0.05**; Supplementary Table 2**). These metabolites reflect habitual dietary intakes within the multi-ethnic study population and are primarily from xenobiotics, lipid and amino acid pathways.

We observed multiple correlations for Asian ethnic-relevant foods, processed foods, fruits and vegetables, red meat, seafood, and nuts and seeds (FDR P<0.05; **Supplementary Table 2**). Chapati, idli, thosai and rasam soup, which are frequently consumed by Indians, were uniquely correlated with betainised compounds (homostachydrine and tryptophan betaine) and other metabolites (X-11849, X-11858, X-21661, X-11847 and X-23639; **Fig. 2**). 4-Hydroxychlorothalonil, a metabolite of crop fungicide, Chlorothalonil, widely used to manage fungal diseases in agricultural industries[17], was correlated with plant-based foods including vegetables, legumes, nuts and seeds. Distinct correlations were observed for roti (X-21383), apple (branched chain 14:0 dicarboxylic acid, methyl glucopyranoside, X-11315 and X-24475) and tropical fruits including melon (X-25271), pineapple (4-allylphenol sulfate) and papaya (N2,N5-diacetylornithine and N-delta-acetylornithine). N2,N5-diacetylornithine was previously reported elevated in blood after intakes of legumes[18] and dry-bean enriched[19] diets. Phytate, derived through ruminant metabolism of chlorophyl[20], correlated with dairy products. Nuts and seeds correlated with 2-hydroxydecanoate, 1-margaroyl-2-linoleoyl-GPC and amino fatty acids (N-acetyl-2-aminooctanoate and N-acetyl-2-aminooctanoate). X-17685, X-13728 and X-11795 correlated with tea, chocolate-containing products and mushroom, respectively.

We identified multiple metabolic associations for animal-based foods. X-11381 correlated with burger and red meat (r=0.20-0.25), similar to previous reports with red meat and milk[21]. Anserine and anserine/carnosine ratio were previously reported as discriminant urinary biomarkers of poultry and white/red meat intake[22,23], but neither were detected in plasma in our study, highlighting i) the lack-of and importance of biomarker validation and relevance in non-European and free-living populations, and ii) dietary biomarkers may vary with biological matrix (urine vs blood), since urine reflects short-term excretion due to rapid turnover while blood reflects circulating metabolites with mid- to long-term relevance. Specific lipid classes were characteristic for red meat, poultry and egg including phosphatidylethanolamines and plasmalogens containing stearoyl moiety, and were distinct from the sphingomyelins, phosphatidylcholines, lysophospholipids and fatty acids characteristic for dairy products (**Fig. 3; Supplementary Table 2**). Tridecenedioate and sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0) were

previously associated with cheese[21]. Deep fried foods correlated with fatty acids, 2-hydroxylaurate and eicosanedioate, and other metabolites (X-11372 and X-21339).

We also observed multiple known biomarkers of fish, coffee, cruciferous vegetables, mushroom and orange in our study, validating their relevance in our multi-ethnic Asian population and in plasma matrix. Fish had the most correlations (39 metabolites), where 3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)[24,25] and hydroxy-CMPF[26] were most strongly correlated, followed by various n3 and n6 PUFA[25], lysophospholipids and plasmalogens containing docosahexaenoic acid (DHA) or eicosapentaenoic acid (EPA) moiety (**Supplementary Table 2**). Additional fish-related correlations included X-13866, X-25417, X-23974 and X-23782. Seafood is generally categorised as a broad food group but in this study, was categorised separately to reflect different organisms. Distinct correlations for crustacean and mollusk were observed for X-23587, X-25810 and X-25419. Coffee had the next most correlations (27 metabolites) with caffeine, xanthines, alkaloids (trigonelline), phenolic acids, diketopiperazines (cyclo(leu-pro)), reported in global population studies[12,16,27] and X-23639.

Other known biomarkers replicated in our study population included ergothioneine and its metabolite, hercynine for mushroom[28,29]; sulfur-containing amino acids (S-methylcysteine sulfoxide, S-methylcysteine and S-methylmethionine) for cruciferous vegetables[21,30]; proline betaine, N-methylproline and N-methylhydroxyproline for orange/citrus fruits[31-34]; β-cryptoxanthin for orange, pineapple and papaya[35,36]; 2-aminophenol sulfate (microbial metabolite of benzoxazinoids found in wheat and rye[37]) for wholegrain foods; 3-bromo-5-chloro-2,6-dihydroxybenzoic acid, 3,5-dichloro-2,6-dihydroxybenzoic acid and N,N,N-trimethyl-5-aminovalerate for dairy/milk[21,25] (**Supplementary Table 2**).

Biomarker panels comprise shared and distinct metabolites

To demonstrate the potential utility for plasma metabolites as objective assessment of dietary intakes in our Asian epidemiological setting, we derived candidate dietary multi-biomarker panels for key foods and beverages, in training datasets (70% of study population), using elastic net penalisation as a machine-learning feature selection approach. Selected metabolites in these biomarker panels were applied to respective test datasets (30% of study population) to evaluate model performance and validation. The biomarker panels were curated for broader food/beverage groups (e.g., fruits), complex mixed dishes (e.g., noodle dishes) and specific foods (e.g., orange) based on questionnaire data, and achieving model variance adjusted R^2>0.1 (to minimise less precise model predictions).

These panels each consisted of 3-39 metabolites (280 total metabolites, of which 211 were unique), primarily from lipid (30%), xenobiotics (16%) and amino acid (16%) pathways (**Fig. 4a**). These metabolites were present after 12 hours of fasting and relevant as mid- to long-term intake biomarkers in the context of diet epidemiology, being less affected by rapid absorption, short elimination kinetics or sampling time window. Sensitivity analyses show these diet-metabolite relationships are not sensitive to genetic ancestry (**Supplementary Table 3**). The panels comprised a combination of i) known diet-derived metabolites (i.e., *CMPF for fish, proline betaine for orange*), ii) metabolites incorporated into endogenous metabolism (i.e., *1-(1-enyl-stearoyl)-2-arachidonoyl-GPE (P-18:0/20:4) for red meat, 1-palmityl-2-arachidonoyl-GPC (O-16:0/20:4) for processed*

meat), iii) gut microbial metabolites (i.e., *2,6-dihydroxybenzoic acid and deoxycholic acid derivatives for fiber-rich foods, anthranilate for noodles*), as well as 51 structurally unidentified metabolites that were iii) previously reported (4%) or iv) potentially novel (22%, **Fig. 4b**). We used a combination of metabolite and genetic approaches to further describe these unidentified metabolites (**Supplementary Table 4**).

Metabolic variation across different foods and beverages showed unique and shared networks (**Fig. 4c**). Notably, metabolite panels for mollusk, poultry and red meat were partially composed of per- and polyfluoroalkyl substances (PFAS) including perfluorooctanesulfonate (PFOS) and perfluorooctanoate (PFOA), persistent environmental contaminants recognised as 'forever chemicals'[38]. Red meat, processed meat, poultry and deep fried foods shared highly correlated networks, including 2R,3R-dihydroxybutyrate (formed via degradation of di/polysaccharides during cooking), X-11299 (associated with processed meat[21]), X-11880, X-16935, X21383 (associated with french fries and/or chips in US and UK populations[21,39]). Ribitol, orotidine and a sulfated piperine metabolite (alkaloid in black/long pepper[40]) were distinct to processed meat panel. Margaroylcarnitine, various plasmalogens, X-12822 and X-22776 were distinct to red meat panel, while various peptides and lipids, X-11308, X-11372, X-21339 and X-23680 were unique to deep fried panel.

Noodle dishes, a complex mixed food group, typically composed of noodles with minor heterogenous proportions of seafood, animal and/or plant-based foods, reflect its metabolite panel, being largest (39 metabolites) amongst the foods, which distinctly include a glucuronidated piperine metabolite, S-allycysteine (biomarker of garlic[41]), androgenic steroids (reflective of animal organs intake), delta-CEHC (metabolite of tocopherol, a fortified ingredient in Asian wheat noodles[42]), anthranilate, X-13507, X-235937, X-11843, X-26106 and X-21310, as well as shared biomarkers of mushroom and seafood.

Besides sharing caffeine metabolites with coffee, tea panel distinctly comprised N-lactoyl valine (formed from free amino acids and lactic acid in fermented foods[43], and also associated with dairy), thymol sulfate, 5-hydroxyindole sulfate, X-17685 and X-23654. Tea and idli/thosai intakes were highly correlated (r=0.68), reflecting Indian-ethnic dietary habits and shared networks via X-13658, homostachydrine and 2,6-dihydroxybenzoic acid, of which the latter two were also present in wholegrain panel. Various dicarboxylate-, polyunsaturated-, long- and medium-chain fatty acids, amino acids, X-17654, X-24494, X-14939, X18799, X-21467 and X11478 were unique to wholegrain panel. Metabolite panels were derived for fruits as a broad category of 17 fruits to capture common compounds (e.g., X-24475 and 4-allylphenol sulfate, a non-specific microbial metabolite of polyphenols[44]), and specifically for orange and papaya, which had strong discriminating metabolites (proline betaine for orange; X-25271 for papaya).

Plant-based metabolites reflect healthy diet quality

As we tend to consume complex meals comprising diverse foods, we assessed diet quality based on internationally established indices (alternative Mediterranean Diet (aMED)[45], alternative Healthy Eating Index (aHEI)[46], Dietary Approaches to Stop Hypertension (DASH)[47], Global Diet Quality Score (GDQS)[48], Plant-based Diet Index (PDI)[49], healthy PDI (hPDI)[49], unhealthy PDI (uPDI)[49]). These dietary indices capture

varied dietary patterns, including plant-based diet, urbanised diet and Mediterranean-style diet, which share similarities to conventional Asian diets rich in vegetables, fish and healthy fats. We derived multi-biomarker panels for these 7 indices, in respective training datasets for each index.

The panels each comprised 30-89 metabolites (total 201 unique), primarily from lipid (31%), uncharacterised (23%), amino acid (18%) and xenobiotics (14%) pathways (**Fig. 4d**). 10 metabolites representing known biomarkers of plant-based foods (beta-cryptoxanthin, ergothioneine, S-methylcysteine sulfoxide, methyl glucopyranoside, 4-hydroxychlorothalonil, carotene diol (1), carotene diol (2), dodecadienoate (12:2), tryptophan betaine, 2,6-dihydroxybenzoic acid) and 2 uncharacterised metabolites (X-11315 and X-24475) were consistently selected across the 6 healthy indices (**Fig. 4e**). Vitamin A metabolism was enriched in all 6 healthy indices (P=2.23x10^{-308}-4.8x10^{-2}; **Supplementary Table 5**), particularly from lipid-soluble isoprenoid pigments abundantly present in various yellow to red-coloured fruits and vegetables. These metabolites mirrored key components of the composite indices (i.e., fruits, vegetables, legumes, wholegrains; **Extended Data Fig. 1**), highlighting their potential as core metabolites of diet quality panels in multi-ethnic and Asian populations.

There were some metabolites that were specific to each of the health indices. Metabolites of coffee (quinate, X-23639), chapati, idli/thosai (homostachydrine), and plant-based oils (GPCs containing linoleoyl and oleoyl moieties) were uniquely present across the PDI indices, while 4-vinylcatechol sulfate, 4-ethylphenylsulfate, X-11299 (metabolite panel of deep fried foods), intermediates of amino acid catabolism and energy metabolism (citrate, 2-aminoadipate, beta-citrylglutamate) and tiglycarnitine were unique to GDQS. In contrast, the multi-biomarker panel of uPDI, suggestive of a poorer diet quality, was distinct in its composition of complex lipids (phosphatidylcholines, lysoplasmalogens, sphingolipids), 3beta-hydroxy-5-cholestenoate (causally linked to atherosclerosis in our study population[50]), hydroxy fatty acids (2-hydroxypalmitate, 2-hydroxyarachidate; involved in complex lipid metabolism), p-cresol glucuronide (microbial metabolite of tyrosine fermentation), and uncharacterised metabolites that were also present in the biomarker panels of deep fried foods and processed meat (X-11308, X-11372, X-13728, X-16935, X-11880, X-21383, X-21319, X-23662). The combination of lauric acid and enriched xanthine metabolism (P=2.23x10^{-308}) suggest higher intakes of coffee sweetened with condensed milk alternatives or creamers, which are commonly formulated with palm kernel oil[51] (rich in lauric acid, ~48%[52,53]) in Asian countries, highlighting attention to population-specific nuances when validating biomarkers.

Biomarker panels improve assessment of dietary intakes

We evaluated effectiveness of the multi-biomarker panels in estimating dietary intakes in respective test datasets, adjusted for age, sex, ethnicity and batch (model 1), additionally for income and years of education (model 2), single biomarker (model 3) and multi-biomarker panels (model 4). Models with multi-biomarker panels explained 3-48% and 0.5-6.5% more variance of intakes, compared to sociodemographics (P=2.23x10^{-308}-1.27x10^{-14}) and single biomarkers (P=7.02x10^{-42}-3.98x10^{-4}, **Fig. 5a**), demonstrating that combination of metabolites offers more specificity and effectiveness than single metabolites, particularly for foods lacking highly discriminant biomarkers.

Coffee panel achieved the most variance (51%, P=2.23x10^{-308}), followed by idli/thosai panel (50%, P=2.12x10^{-43}) and chapati panel (43%, P=5.15x10^{-40}), compared against models 2. Idli/thosai and chapati strongly reflected ethnic-specific foods, and in combination with legumes, are frequently consumed by Indian, compared to Chinese and Malay populations. This finding is consistent with (Indian) ethnicity being a major contributor to the variance for these foods (idli/thosai: P=2.12x10^{-120}; chapati: P=3.17x10^{-78}; legumes: P=6.96x10^{-32}).

Estimated dietary intakes, derived from these biomarker panels showed moderate to good agreement and ranking validity with self-reports, r=0.33-0.72 (**Fig. 5b**). Error was greatest amongst individuals reporting zero intake and excluding this group of individuals improved median prediction accuracy (**Extended Data Fig. 2**). Accuracy was high (>70%) for coffee, tea, red meat, fish, fruits, cruciferous vegetables, deep fried foods and noodle dishes, compared to other foods, demonstrating the potential utility of biomarker panels for objective assessment of dietary intakes. These metabolite panels offer a stepwise approach from qualitative (zero vs presence of intake) to quantitative (low to high intake) assessment.

Biomarker panels for diet quality

Metabolite panels were also used to construct diet quality biomarker models, explaining 7-15% higher variances, compared to models including sociodemographics only (P=1.08x10-^{199}-6.50x10^{-92}, **Fig. 5c**), and showed moderate to good agreement with diet quality scores derived from self-reports (r=0.36-0.58, **Fig. 5d**). Population mean scores ranged from 2.97±1.5 of 9 (aMED), 40.43±9.23 of 90 (aHEI), 20.59±4.83 of 40 (DASH), 19.99±4.66 of 48.5 (GDQS), 45.48±6.61 of 48.5 (PDI), 47.40±7.40 of 48.5 (hPDI) and 47.60±6.07 of 48.5 (uPDI), where higher scores represent healthier diet quality and associated with higher proportion of energy intake from fiber (P=2.23x10^{-308}-9.45x10^{-148}) and healthy fats (PUFA: P=2.23x10^{-308}-6.36x10^{-31}; MUFA: P=7.10x10^{-219}-2.45x10^{-2}) and age (P=8.79x10^{-179}-7.52x10^{-38}; **Extended Data Fig. 1**, **Supplementary Table 6**).
The diet quality scores, derived from plasma metabolite levels, represents an alternative scoring (≥68% prediction accuracy, **Extended Data Fig. 2**) to the conventional approach of aggregating scores (based on self-reports) of individual food/nutrient components into a composite score, which can lead to compounded error and are influenced by scoring quintiles (for DASH, aHEI) determined by study sample size used.

Diet-metabolite relationships are reproducible

We evaluated the temporal consistency of diet-metabolite relationships in a subset of individuals who returned for a repeat visit (**Extended Data Table 2**). Plasma metabolites of the dietary biomarker panels from the baseline visit were present also in plasma profiles at the repeat visit, showing stability of these metabolites and diet-associations over time. Using models of biomarker panels trained in the baseline visit, we estimated intakes of respective foods, beverages and diet quality at the repeat visit, which showed good agreement with self-reports recorded at the repeat visit, r=0.35-0.73 (**Fig. 6a**).

Across visits, the reliability and reproducibility of estimated intakes (intraclass correlation coefficients (ICC)=0.54-0.91) consistently outperformed self-reports (ICC=0.32-0.85), with 50% of foods and beverages achieving ICC>0.75, indicating good

reliability (**Fig. 6b**). Estimated and reported diet quality scores showed similar reliability (ICC=0.62-0.79 vs self-reports: ICC=0.57-0.76).

Metabolite scores link diet to health outcomes

We developed metabolite scores for the 27 foods, beverages and diet quality indices, as a weighted sum of metabolites from the dietary biomarker panels, based on the significantly associated metabolites from self-reports. Compared to self-reports, the metabolite scores were stronger predictors of clinical phenotypes including homeostatic model assessment for insulin resistance (HOMA-IR), type 2 diabetes (T2D), BMI, fat mass index (FMI), carotid intima-media thickness (cIMT) and hypertension (**Fig. 7**, **Supplementary Tables 7 and 8**).

Scores of 'healthy' plant-based foods including fruits, vegetables, wholegrain foods and chapati as well as fish and dairy products were negatively associated with these phenotypes, with the third tertile corresponding to the strongest associations (P=2.39x10^{-119}-2.37x10^{-2}), as expected and consistent with direction of associations based on self-reports (P=3.03x10^{-16}-5.91x10^{-3}). For glycemic and adiposity phenotypes, fruits scores showed most protective effect (HOMA-IR: P=4.34x10^{-110}; BMI: P=7.21x10^{-104}; FMI: P=2.39x10^{-119}; T2D: P=1.78x10^{-45}). For vascular disease risks, legume and fruits scores were most protective, respectively (cIMT: P=2.29x10^{-18}; hypertension: P=5.62x10^{-17}). Scores of healthy diet quality indices showed similar trends where aMED was most protective for FMI (P=6.43x10^{-199}), BMI (P=7.183x10^{-176}), and cIMT (P=9.18x10^{-25}), while PDI was most protective for HOMA-IR (P=2.35x10^{-122}), T2D (P=8.83x10^{-55}) and hypertension (P=8.81x10^{-23}), respectively.

In contrast, other seafood, animal-based and processed foods, and uPDI associated positively with these cardiometabolic phenotypes (scores: P=1.05x10^{-61}-3.08x10^{-2}; self-reports: P=6.67x10^{-19}-3.61x10^{-2}). Environmental contaminants, PFAS, that comprise several biomarker panels, also show positive associations with all phenotypes (P=2.42x10^{-57}-4.06x10^{-2}; **Extended Data Fig. 3a**). Amongst the 'unhealthy' foods, processed meat scores showed the most detrimental effect (HOMA-IR: P=6.53x10^{-60}; BMI: P=1.05x10^{-61}; FMI: P=1.69x10^{-45}; cIMT: P=7.22x10^{-18}; T2D: P=6.27x10^{-24}; hypertension: P=2.38x10^{-21}). Red meat score associated negatively with most phenotypes (P=2.23x10-^{17}-1.91x10^{-2}), contrasting the positive associations from self-reports in our study (P=2.29x10^{-6}-2.38x10^{-2}) and meta-analysis of three US cohorts[54], suggesting preparation methods[55] and biomarker composition play a role in influencing health associations. The red meat panel comprised primarily PUFA, and lesser proportions of amino acids, and *trans* and saturated fatty acids ('unhealthy fats'), likely contributing to overall negative associations. Metabolite scores for noodles dishes showed higher risks with all phenotypes (P=3.86x10^{-29}-1.59x10^{-4}), consistent with its panel comprising ceramides, where elevated levels have been linked to insulin resistance, obesity and CVD[56].

Coffee and tea intakes, derived from both scores (P=2.98x10^{-18}-3.97x10^{-2}) and self-reports (P=1.43x10^{-8}-3.87x10^{-2}), were positively associated with most phenotypes, considering the trend of intakes in this study population being predominately (>55%) energy-dense variations (with condensed/evaporated/whole milk, sugar or 3-in-1, **Extended Data Fig. 3b**). Overall, the associations with metabolite scores were robust to BMI (M2) as a mediator, unlike the associations with self-reports, which were no longer

significant, and most associations remain consistent and significant after adjusting for socioeconomic variables (income, education), smoking status and physical activity levels (M3; **Fig. 7**, **Supplementary Tables 7 and 8**).

Discussion

We generated fasting plasma profiles of 1,055 metabolites in 8,391 individuals (195 with repeat data) for whom we had extensive dietary records, sociodemographic, health and lifestyle variables. We carried out diet-metabolite association analyses and derived 27 dietary biomarker panels that were representative of specific food/beverage intakes and complex meals within a multi-ethnic Asian population, using machine-learning based prediction models. We showed the feasibility and utility of these biomarker panels to objectively assess dietary intakes and quality, compared to self-reports, whose limitations are well-documented.

Biomarker-calibrated models for single foods[15,57,58] based on highly specific biomarkers and broader metabolomics profiles characterising diet quality (DASH[58] and HEI[59]) have been reported in American and European populations but not developed or validated in other populations. Considering the chemical complexity of diets in different regions, our study highlights the importance of mapping the highly individualised Asian metabolome in a geographically and culturally relevant context. When used to assess daily intakes, some of these biomarker panels captured ethnic differences in dietary habits (e.g., idli/thosai and noodle dishes), which is key to documenting diets underlining a multi-ethnic population. Metabolite-based modelling represents an avenue to explore, particularly for under-studied culturally-relevant foods, compared to common/Westernised foods.

While some foods have uniquely discriminant biomarkers (e.g., proline betaine for orange), our findings demonstrate that using a combination of metabolites (as a panel) consistently explained higher variance of dietary intakes and significantly improved intake estimations, compared to single biomarkers. We observed metabolites across the panels, reflecting foods with overlapping chemical composition, shared agricultural practices or convergent metabolic responses. Some metabolites may represent true biomarkers of intake, being derived directly from dietary or environmental exposures, while others reflect metabolic responses to diet intake, from endogenous biochemical pathways or gut microbial metabolism. For instance, amongst the endogenous metabolites, several lipid species (ceramides, sphingomyelins, plasmalogens) are primarily synthesised de novo or through distinct metabolic pathways (sphingomyelin hydrolysis, sphingomyelinase-mediated activity, ether lipid biosynthesis) with essential roles in cell membrane integrity and function, cell signalling and oxidative stress[56,60], while most amino acids, either synthesised de novo or formed as pathway intermediates, are integral to protein synthesis, neurotransmission, energy homeostasis and regulate processes such as bile salt conjugation[61]. While differentiating between metabolites of direct dietary exposure or those arising from metabolic adaption to diet introduces complexity in interpretation, this does not diminish their relevance and potential utility as a biomarker of dietary intake or pattern.

Dietary intakes estimated by an individual's plasma biomarker panel correlated with self-reports for specific foods, complex dishes and diet quality indices, and these food-metabolite relationships were stable over time. Beyond serving as an intake

prediction tool, the panels behaved as a proxy of intakes or quality that improved accuracy of clinical associations, reinforcing their potential as more reliable and reproducible biomarkers of dietary exposure. The identified metabolites have potential utility in clinical settings and nutrition research. In controlled feeding or intervention studies, these metabolites or metabolic variation could be used to monitor dietary adherence and compliance (reducing reliance on self-reports and participant burden[62]), or disease risk prediction[63]. These metabolites may also facilitate personalised nutrition strategies by characterising (individual or metabotype) metabolic responses to dietary habits for tailored dietary recommendations to mitigate specific diet-related health outcomes.

We acknowledge our study is observational with most individuals having dietary and metabolite datasets at one timepoint and the metabolite-phenotype associations do not highlight causal relationships, a similar limitation encountered by other epidemiological studies. Plasma metabolite levels mirror recent diet history and may not reflect long-term kinetics beyond the metabolites half-times (few hours to days) depending on frequency of intakes. Some metabolites within the panels are structurally unidentified and hence not well-studied in the context of their derived source or chemistry, and should be interpretated with caution. Our metabolite associations and panels, particularly for ethnic-specific foods will benefit from replication in other Asian populations.

Our study provides insights into diet-related metabolic variations, including ethnic- and region-specific dietary habits, and environmental effects on the food supply. The HELIOS study population closely mirrors the ethnic and sociodemographic distribution of the adult Singapore population[64], enhancing generalisability of our findings to multi-ethnic urban populations, in contrast to relatively homogenous population studies. We demonstrate that the utility of dietary biomarker panels, objectively representative of food and beverage intakes in complex diets, is feasible, and generates clinically-relevant health outcomes in Asian populations.

Methods

General cohort information. Datasets were derived from the Health for Life in Singapore (HELIOS) study (ethics approval by Nanyang Technological University Institutional Review Broard: IRB-2016-11-030), a multi-ethnic prospective cohort of adults aged 30-84 years old[64]. In HELIOS, 10,004 individuals (Singapore citizens or permanent residents) were recruited from the general population across socio-economic backgrounds, and excluded if they were pregnant, breastfeeding, had prior major illness requiring hospitalisation/surgery, cancer treatment in the past year or recent participation in drug trials. Individuals provided written informed consent and underwent an extensive standardised assessment protocol that included i) questionnaires, ii) physiological measurements, iii) imaging procedures, iv) biological sample phenotyping, and were compensated $50 to cover travel- and participation-related expenses.

Individuals' age, sex and ethnicity were recorded as per their identity card. Environmental and lifestyle information including income level, years of education, cigarette smoking and medical history were recorded using self-administered questionnaires. Income level was categorised into "<S$4,000", "S$4,000-9,999" and "≥S$10,000". Smoking status was categorised as "non-smoker", "ex-smoker" or "smoker". In this study, we excluded individuals not of primary ethnicities (Chinese, Malay, Indian) or who did not complete physiological or biochemical assessments, Food Frequency

Questionnaire (FFQ) or reported implausible total energy intake, or who did not have metabolomics data (**Fig. 1**). Physical activity (n=7,583 individuals) in various domains and intensity levels were derived from the long iPAQ. Participants with ≤150-300 min of moderate intensity or ≤75-150 min of vigorous activity or an equivalent of moderate- and vigorous-intensity activity throughout the week and with 2 days on moderate- and vigorous intensity activities/week were considered to meet WHO recommendations, and excluded if they had >960 min/day.

Clinical phenotyping. Fasted blood biochemistry including fasting glucose, HbA1c, cholesterol (total, LDL, HDL) and triglycerides were measured by a local accredited laboratory (QuestLab Singapore, SAC-SINGLAS ISO 15189:2012). Insulin was measured in-house using ADVIA Centaur XP Immunoassay System (Siemens Healthcare, Erlangen, Germany). HOMA-IR was calculated as insulin (mIU/L)*glucose (mM)/22.1512. FMI was derived from body fat mass (kg) quantified using whole-body Dual X-ray Absorptiometry (DEXA) scans (Horizon^{TM} W densitometer S/N 30052M, Hologic, Massachusetts, USA) and normalised against height (m^2). cIMT, a subclinical atherosclerotic phenotype, was estimated from carotid ultrasound imaging[50] of the common carotid artery (Philips iU 22 ultrasound system) and mean cIMT values were natural log-transformed for downstream analyses. Individuals were classified with T2D based on ≥7 mM fasting glucose level, ≥6.5% HbA1C, prior diagnosis by doctor (self-reported history) or taking diabetes medications, based on American Diabetes Association[65] and WHO[66] recommendations. Individuals were classified with hypertension based on systolic blood pressure (SBP) ≥140 mmHg, diastolic blood pressure (DBP) ≥90 mmHg, prior diagnosis by doctor (self-reported history) or taking hypertensive medications, based on International Society of Hypertension recommendations[67].

Reproducibility study. There were 235 individuals for whom data and samples had been collected from baseline and repeat visits, 336 days apart on average, to demonstrate reproducibility of research phenotype characterisation (**Extended Data Table 2**). Individuals completed the same questionnaires and clinical phenotyping across baseline and repeat visits. These individuals were used to evaluate the temporal consistency of diet and metabolite relationships described below.

Data generation and preprocessing

Dietary assessment. Dietary intakes were assessed by a self-reported semi-quantitative validated[68] FFQ, comprising 169 food and beverage items. Individuals reported the serving size and frequency of food/beverages consumed daily (1, 2-3, 4+ times/day), weekly (1, 2-3, 4-6 times/week) or monthly (1, 2-3 times/month). Intakes were then converted to grams/day by multiplying frequency of intake by portion size. Portion sizes were assigned a specified grams amount based on the Singapore Health Promotion Board composition database (*https://focos.hpb.gov.sg/eservices/ENCF*). Daily nutrient (grams/day) and energy (kcal/day) intakes were also calculated using the same database. Individuals were excluded if their energy intakes were considered implausible based recommended energy intake thresholds, i.e., <500 kcal/day or >3 SD of gender specific mean)[69]. For downstream analysis, food/beverage items were aggregated into 42 broader

groups based on biological classification and/or nutritional property (**Extended Data Table 3**), with total intake (grams/day) for each group calculated as the sum of its constituent foods/beverages. Individuals with intakes (grams/day) >99.5 percentile were considered outliers and excluded.

Diet quality was assessed based on internationally established indices (DASH[70], aMED[71], aHEI[72], GDQS[48] and PDI[73]; **Extended Data Fig. 1**). These indices were selected for their global validity across diverse populations, cultural adaptability (can be modified to include regionally relevant foods) and relevance to diet-related chronic diseases prevalent in Asian populations including Singapore[74] (CVD, T2D, hypertension). DASH and aHEI scores were calculated for 7,742 individuals for whom sodium intakes were available. DASH[70] score ranged from 8-40, where higher score indicates a diet associated with lower blood pressure and reduced CVD risk. aMED[71] score ranged from 0 to 9, with higher score indicating higher adherence to a Mediterranean-style diet pattern, widely recognised as a healthy diet. The modified aHEI[72] (no trans-fat, EPA+DHA) score ranged from 0 to 90, with higher score indicating healthier diet patterns associated with lower risk of chronic diseases such as CVD, T2D and certain cancers. The modified GDQS[48] (no deep orange tubers) score ranged from 0 to 48.5, with scores ≥23 associated with low risks of nutrient adequacy and NCD. There are 3 graded versions of PDI[73] to emphasise intakes of plant-based foods while reducing animal-based food intakes, and score ranged from 18-90. In UPDI, negative scores are given to unhealthy plant-based foods such as refined grains and sugar, and processed foods.

Plasma metabolome. Untargeted metabolomics profiling of fasting EDTA plasma samples was performed by Metabolon, Inc (North Carolina, USA)[75-77]. Samples (n=9,995) comprised of individuals with baseline (n=9,760) and repeat (n=235) visits, which were initially stored at -80°C, thawed, aliquoted and shipped on dry ice to Metabolon. Briefly, samples were methanol-extracted and analysed by untargeted liquid chromatography-tandem mass spectrometry (LC-MS/MS) Metabolon Global Discovery platform using 4 methods with diverse column and gradient conditions (2 reversed-phase (RP) methods in positive electrospray ionisation (ESI) mode, a RP method in negative ESI and a HILIC method in negative ESI), from 70-1000 m/z range. A cocktail of quality control (QC) standards was added to each sample (before injection) for monitoring of instrument performance and chromatographic alignment. Aliquots from each sample were pooled to generate a pooled matrix sample that was periodically injected as technical replicates and to distinguish between biological and process variability. Samples (collectively from baseline and repeat visits) were randomised within batch runs, across 5 batches. Raw peak area-under-the-curve values (relative intensity of ion counts) were quantified, normalised within each analytical batch to account for inter-day instrumental variability using median-scaling, and further normalised across the 5 analytical batches to account for batch-to-batch variability using median-scaling.

1,696 metabolites were initially identified. Confidence levels for identification of metabolites were assigned based on Metabolomics Standards Initiative (MSI)[78] and metabolites were categorised to nine pathways corresponding to their chemical class. Metabolites reported with single (*) or double (**) asterisks were not confirmed against a known chemical standard/structure and refer to MSI level 2 or 3. Metabolites of unknown structural identity were assigned numeric reference numbers beginning with X. Samples

(n=5) that failed QC standards were excluded. Median relative standard deviation (SD) of instrument and technical variability between batch runs ranged from 6-9% and 10-12%, respectively. Metabolites with >20% missingness (assuming not detected) were excluded and the remaining were imputed with minimum value for each metabolite (assuming missingness due to low abundance). Metabolites were further excluded if they were considered outliers (defined as >6 SD away from the mean of the first 2 principal components in overall and method-specific Principal Component Analysis, leading to 1,055 metabolites.

Statistical analysis

Metabolome-wide correlation with dietary habits. Dietary intakes and plasma metabolite levels were natural log-transformed and scaled for downstream parametric analyses to improve normality and variance homogeneity. Assumptions of normality and equal variance were assessed using formal tests and inspection of residuals. To explore the relationship between diet and metabolites, we performed a discovery analysis using partial correlation analysis of metabolites against 169 foods/beverages and 42 broader food/beverage groups, adjusted for age, sex, ethnicity and batch effects (*pcor* v1.1). *P* values were adjusted for multiple testing (1055 pairwise correlations between metabolites) with Benjamin-Hochberg false discovery rate (FDR) method. To select statistically and biologically meaningful associations, only significant correlations (FDR $P<0.05$) with spearman rank correlation coefficients, $r \geq 0.15$ were considered (**Supplementary Table 2**).

Estimating variance of individual metabolites. To estimate the variance of each metabolite that contributed to each food/beverage group, we applied univariate OLS regression (*glmnet* v4.1-7), adjusted for age, sex, ethnicity and batch. *P* values were FDR-adjusted and considered significant if FDR $P<0.05$.

Nutrient composition across diet quality indices. Daily nutrient (grams/day) intakes were converted to proportion of energy intakes (% kcal/day) based on 4 kcal/gram (sugar, starch and fibre), 4 kcal/gram (protein) and 9 kcal/gram (saturated and unsaturated fats). To estimate the effect size of each nutrient across each diet quality index, we applied univariate regression, adjusted for age, sex, ethnicity and energy intake. *P* values were FDR-adjusted and considered significant if FDR $P<0.05$ (**Supplementary Table 6**).

Metabolite panels for assessing dietary intakes. We examined the associations of metabolites with dietary intakes in separate discovery and validation phases, by partitioning the study population into training (70%) and test (30%) datasets, to mitigate overfitting and allow independent validation of findings. Stratified sampling was performed to ensure training and test datasets contained balanced proportions of age, sex and ethnicity (refer to **Supplementary Table 9** for training and test datasets' sample sizes).

As a variable selection method and for handling highly correlated predictors, we applied machine learning-based elastic net[79] regression with 10-fold cross-validation to each training dataset. For each foods/beverage model, covariates (age, sex, ethnicity, batch, income and years of education) were included in elastic net regression models without penalty factor (implying no shrinkage), unlike the metabolites whose coefficients

were subjected to lambda, λ, penalty term. For each diet quality index, energy intake was additionally included as a covariate, to adjust for energy-density bias and energy-adjusted nutrient components across the indices. Optimal alpha, α, values were tuned at 2 levels, from 0 to 1 (resolution=0.1) and subsequently from ±0.05 of α (at the minimum mean-squared error, resolution=0.01). To achieve a parsimonious multi-biomarker panel for each food/beverage (outcome variable), metabolites derived from each model with non-zero positive coefficients were inspected and excluded from respective panels based on condition that they were not more strongly associated with other food/beverage items.

Each food/beverage intake was then assessed, in respective test datasets, using independent regression models, adjusted for age, sex, ethnicity and batch (model 1), additionally for income and years of education (model 2), single biomarker (corresponding to the metabolite with most significant FDR *P* and largest effect size within the metabolite panel; model 3) and multi-biomarker panels (model 4). Based on literature and chemical composition, we recognise that not all foods/beverages will have highly discriminant biomarkers. To minimise less precise model predictions, we retained 16 food/beverage groups whose metabolite panels have reasonable explanatory power (adjusted R^2>0.1), adjusted for age, sex, ethnicity and batch. Additional models were proposed for 4 specific foods including orange, papaya, chapati and idli/thosai as these items exhibited highly specific biomarkers (univariate FDR P=1.28x10^{-294}-1.34x10^{-34}) and/or are uniquely consumed in our multi-ethnic Asian population.

Metabolite panels for assessing diet quality. For diet quality indices, a similar approach was applied for partitioning into training (70%) and test (30%) datasets (refer to **Supplementary Table 7** for training and test datasets' sample sizes), and variable selection of metabolites, adjusting additionally for energy intake in the elastic net regression models. The diet quality indices reflect overall diet patterns composed of multiple foods and beverages, hence, all metabolites derived from each model with non-zero positive coefficients were included in the metabolite panels to capture the composite nature of diet patterns. Each diet quality index score was then assessed in respective test datasets, using independent regression models, adjusted for age, sex, ethnicity and batch (model 1), additionally for income and years of education (model 2) and multi-biomarker panels (model 4).

Diet-metabolite networks. Partial correlations within and between metabolites (of the metabolite panels) and foods/beverages were summarised into graphical networks using fruchterman-reingold layout (*ggraph* v2.1.0).

Agreement between reported and estimated dietary intakes. Metabolite panels were evaluated for their effectiveness as a tool for estimating food/beverage intakes or diet quality, in the test datasets (baseline visit), using spearman rank correlation coefficient, *r*, for agreement between reported and estimated intakes or diet quality (*stats* v4.2.1).

Diet-based metabolite scores and health outcomes. Composite scores were developed for each multi-biomarker panel as an objective assessment of each food and beverage intake for each individual, by applying penalised weights derived from ridge regression coefficients to the metabolites abundance and taking the sum. The composite

scores were subsequently categorised into tertiles and tested for associations with clinical phenotypes (HOMA-IR, T2D, BMI, FMI, cIMT and hypertension) in comparison to self-reported intakes (also categorised into tertiles) in independent OLS regression models (for HOMA-IR, BMI, FMI, cIMT) or logistic regression models (for T2D and hypertension), adjusted for age, sex, ethnicity and batch. As socioeconomic variables (income, education), smoking status and physical activity levels may confound the associations between diet/metabolites and tested phenotypes, we additionally adjusted for these covariates in separate models. Given the potential role of BMI in the pathophysiology and metabolic regulation of the tested phenotypes, we performed additional sensitivity analysis including BMI as an covariate. Clinical data was scaled to allow comparison across phenotypes. In all cases of multiple testing, *P* values were FDR-adjusted and considered significant if FDR-*P*<0.05.

Temporal consistency of diet-metabolite relationships. Of the 235 individuals from the reproducibility study, 40 individuals with implausible daily energy intakes[69] were excluded, remaining 195 individuals with data from baseline and repeat visits, 322 days apart on average (~11 months). Models, for each food and beverage, trained using data from the baseline visit was used to estimate respective intakes in the repeat visit, and assessed for agreement between reported and estimated intakes via spearman rank correlation coefficient, *r*. The test-retest reliability within reported intakes, estimated intakes and metabolite scores across visits were determined via ICC using a two-way mixed effects model (type= consistency; *irr* v0.84.1).

Characterisation of unidentified metabolites. To better describe the 51 metabolites (with unknown structural identity), a combination of metabolite and genetic approaches were used. **Metabolite-metabolite correlations.** We performed full order partial correlation analysis between all metabolites to assess which unidentified metabolites (unknowns) were structurally/chemically correlated to metabolites of known identity, controlled for effects of age, sex and ethnicity. *P* values were Bonferroni-adjusted. For each unknown, paired coefficients $r \geq 0.1$ with metabolites of known identity were reported in **Supplementary Table 2**. **GWAS and pathway analysis.** Genomic data was processed for quality control and downstream analyses[64]. Briefly, GWAS was performed for each metabolite and ethnic group using linear mixed regression framework implemented in REGENIE[80], adjusted for age, sex and first 20 genetic PCs. The results were meta-analysed using METAL[81] to obtain trans-ancestry summary statistics. Significant independent associations for unknowns were identified by clumping GWAS summary statistics using PLINK --clump function[82] ($P<5X10^{-8}$, LD $r^2<0.1$ and 500 KB window size). The resulting independent metabolite-SNP associations above genome-wide significance ($P<5x10^{-8}$) and genes/loci in the top associations were reported in **Supplementary Table 10**. We queried the GWAS catalog[83] to identify which of our associations have been previously reported and to find other associations (not found in our study) for the corresponding unknowns. To further identify biological processes, molecular functions and/or KEGG pathways enriched for these unknowns, we queried the genetic determinants (from our current study and previously reported in GWAS catalog) in STRING (v12.0)[84]. To enable meaningful interpretation,

these information were used, altogether, to describe these unknowns (**Supplementary Table 4)**.

Data availability

The HELIOS study individual-level datasets are not publicly available due to data privacy regulations. Researchers may apply for access through HELIOS Data Access Committee (helios_science@ntu.edu.sg). Summary-level datasets are available as source data including food-metabolite and metabolite-metabolite partial correlation coefficients, correlation coefficients between self-report against estimated intakes, intraclass correlation coefficients from reproducibility study, and regression coefficients relating self-reports and metabolite scores with clinical phenotypes.

Code availability

Data processing, analysis and visualisation were performed using *R* (v4.5.0) and the following packages: *tidyverse* (v1.3.2), *readxl* (v1.4.1), *glue* (v1.6.2), *pcor* (v1.1), *glmnet* (v4.1-7), *stats* (v4.2.1), *irr* (v0.84.1), *ggplot2* (v3.4.2) and *gplots* (v3.1.3). Codes for analyses and visualisation are available via GitHub at https://github.com/HELIOS-SG100K-LKC/Metabolic-variation-reflects-dietary-intake-in-a-multi-ethnic-Asian-population.

Acknowledgements

We express our thanks to participants of the HELIOS study and the HELIOS operation team for recruitment, organisation and data/sample collection. This study (NTU IRB: 2016-11-030) is supported by Singapore Ministry of Health's (MOH) National Medical Research Council (NMRC) under its OF-LCG funding scheme (MOH-000271-00), Singapore Translational Research (StaR) funding scheme (NMRC/StaR/0028/2017), National Research Foundation (NRF) through NMRC and Precision Health Research (PRECISE) through the National Precision Medicine programme and intramural funding from Nanyang Technological University, Lee Kong Chian School of Medicine and National Healthcare Group.

Author contributions

J.C. J.B., P.E., E.R., J.L., E.S.L. and J.N. acquired funding for HELIOS study. J.C., K.W., P.A.S., R.S. and G.M. acquired funding for metabolomics analysis. J.C. and D.Y.L. conceptualised the study. D.Y.L., T.H.M., N.S., P.R.J., R.D., K.W., P.A.S., R.S. and G.M. generated the data. D.Y.L., T.H.M., N.S. and P.R.J. analysed the data. J.C. and D.Y.L. wrote initial manuscript draft. M.L., T.H.M., N.S., P.R.J. and P.A.S. provided critical feedback to manuscript draft. All authors reviewed and contributed to revision of the manuscript.

Competing Interests

K.E.W and G.A.M are employees of Metabolon. The remaining authors declare no competing interests.

Figure Legends/Captions (for main text figures)

Fig. 1. Overview of study design. The HELIOS study comprises a population-based sample of 10,004 multi-ethnic Asian men and women living in Singapore. This study explored the relationships between habitual dietary exposures (from self-reported qestionnaire) and plasma metabolomics profiles (from untargeted mass-spectrometry based metabolomics). Multi-biomarker panels and composite scores were built using machine learning models to estimate dietary intakes and quality, and tested for associations against cardiometabolic clinical outcomes. Reproducibility of diet-metabolite relationships were assessed in a subset of 195 individuals, with baseline and repeat visits averaging 336 days apart. Parts of figures were created with Biorender.com.

Fig. 2. Metabolome-wide association heatmap (non-lipid metabolites). Heatmap of foods and beverages with non-lipid metabolites (from amino acid, carbohydrate, cofactors and vitamins, peptide and xenobiotics pathways, and structurally unidentified) from partial correlation analysis (two-sided), adjusted for age, sex, ethnicity and batch, within the HELIOS study (n=8,391). *indicates FDR P<0.05 and spearman coefficients (r)≥0.15. Ordering of variables were determined by hierarchical clustering.

Fig. 3. Plasma metabolome-wide association (lipid metabolites). Heatmap with foods and beverages with lipid metabolites (from lipid pathway) from partial correlation analysis (two-sided), adjusted for age, sex, ethnicity and batch, within the HELIOS study (n=8,391). *indicates FDR P<0.05 and spearman coefficients (r)≥0.15. Ordering of variables were determined by hierarchical clustering.

Fig. 4. Summary of dietary metabolite panels for specific foods, beverages and diet quality. Candidate dietary multi-biomarker panels for foods and beverages a) assigned to super pathways, and b) categorised into reported or potentially novel. C) Network graph summarising shared and distinct networks across dietary variables (foods and beverages) and biomarker panels, based on partial correlation analysis (two-sided) adjusted for age, sex and ethnicity. Each node represents a food, beverage or metabolite. Edge colour represents direction of correlation (red for positive, blue for negative) and edge thickness represents strength of correlation. Diet quality multi-biomarker panels showing d) distribution of metabolites by super pathways, and e) composition of metabolites (n=201) across the 7 diet quality indices. Cells with an asterisk represent metabolites contained in each biomarker panel. Metabolite labels in red are common across the 6 healthy indices and labels in blue are unique across each index. Line colour represents direction of correlation (red for spearman r≥0.7, blue for spearman r≤0.3). aMED, alternative Mediterranean Diet; aHEI, alternative Healthy Eating Index; DASH, Dietary Approaches to Stop Hypertension; GDQS, Global Diet Quality Score; hPDI, healthy Plant-based Diet Index; uPDI, unhealthy Plant-based Diet Index.

Fig. 5. Assessment of dietary intakes and quality from dietary metabolite panels. a) Dietary multi-biomarker panels improved estimation of intakes of foods and beverages and b) diet quality scores in independent regression models of respective test datasets (n=2,336-2,513). Exact sample sizes for each of the 27 test datasets are listed in **Supplementary Table 9**). Values represent proportion of variance estimated. Significant difference between models (model 3 vs 4 in Fig. 4a (P value range: 2.23×10^{-203}-$1.27 \times 10^{-}$

792 ^{14}); model 2 vs 4 in Fig. 4b (*P* value range: $2.59x10^{-108}$-$6.21x10^{-30}$)) is represented by
793 ****P*<0.001. c) Scatter plots showing agreement between self-reported (x-axis) vs
794 estimated (y-axis) intakes (natural log-transformed grams/day) for each food, beverage
795 and diet quality index, from two-sided spearman correlation analysis, with 95%
796 confidence levels shown as error bars. Each point represents an individual, coloured by
797 ethnicity.
798
799 **Fig. 6. Reproducibility of diet-metabolite relationships across visits.** a) Scatter plots
800 showing agreement between self-reported (x-axis) vs estimated (y-axis) intakes (natural
801 log-transformed grams/day) for each food and beverage in the repeat visit (n=195 pairs
802 of individuals), from two-sided spearman correlation analysis, with 95% confidence levels
803 shown as error bars. Estimated intakes were derived from models trained on training
804 datasets from baseline visit. Each point represents an individual, coloured by ethnicity. b)
805 Intraclass correlation coefficients (ICC) showing test-retest reliability of reported and
806 estimated intakes across baseline and repeat visits (n=195 pairs). Error bars represent
807 95% confidence intervals.
808
809 **Fig. 7. Associations between dietary intakes, diet quality and clinical phenotypes.**
810 Effect sizes of dietary intakes (third tertile; first tertile as reference) and diet quality,
811 derived from self-reported intake or metabolite score, adjusted for age, sex, ethnicity and
812 batch (M1, n=8,391). Estimates were obtained from independent linear regression (for
813 HOMA-IR, BMI, FMI, cIMT) or logistic regression (for T2D, hypertension) for each food
814 and diet quality, based on following sample sizes: T2D (n=1,040 T2D, n=7,351 non-T2D),
815 hypertension (n=2,411 with hypertension, n=5,980 without hypertension). Association
816 strength represents log-transformed FDR *P* values. Circle border represents significant
817 associations (FDR *P*<0.05). Sensitivity analysis of effect sizes, further adjusted for
818 potential confounders including BMI (M2), and for smoking status, education, income and
819 physcial acitvity levels (M3). HOMA-IR, homeostatic model assessment for insulin
820 resistance; T2D, type 2 diabetes; FMI, fat mass index; cIMT, carotid intima-media
821 thickness.
822

References

1 World Health Organisation. *Cardiovascular diseases (CVDs): Key facts*, <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)> (11 June 2021).

2 Barouki, R., Gluckman, P. D., Grandjean, P., Hanson, M. & Heindel, J. J. Developmental origins of non-communicable disease: Implications for research and public health. *Environmental Health* **11**, 42 (2012). https://doi.org:10.1186/1476-069X-11-42

3 Afshin, A. *et al.* Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **393**, 1958-1972 (2019). https://doi.org:10.1016/S0140-6736(19)30041-8

4 Uusitalo, U., Pietinen, P. & Puska, P. Globalization, Diets and Noncommunicable Diseases. *Dietary transition in developing countries: challenges for chronic disease prevention* (2002). <https://apps.who.int/iris/bitstream/handle/10665/42609/9241590416.pdf>.

5 Xu, Y. Y. *et al.* The Global Burden of Disease attributable to low physical activity and its trends from 1990 to 2019: An analysis of the Global Burden of Disease study. *Frontiers of Public Health* **10**, 1018866 (2022). https://doi.org:10.3389/fpubh.2022.1018866

6 Brennan, L., Hu, F. B. & Sun, Q. Metabolomics meets nutritional epidemiology: harnessing thepotential in metabolomics data. *Metabolites* **11**, 709 (2021). https://doi.org:10.3390/metabo11100709

7 Agudo, A. Measuring intake of fruit and vegetables. (2005). <https://apps.who.int/iris/bitstream/handle/10665/43144/9241592826_eng.pdf>.

8 Scalbert, A. *et al.* The food metabolome: a window over dietary exposure. *American Journal of Clinical Nutrition* **99**, 1286-1308 (2014). https://doi.org:10.3945/ajcn.113.076133

9 Barabási, A.-L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nature Food* **1**, 33-37 (2020). https://doi.org:10.1038/s43016-019-0005-1

10 Landberg, R. *et al.* Future prospects for dissecting inter-individual variability in the absorption, distribution and elimination of plant bioactives of relevance for cardiometabolic endpoints. *Eur J Nutr* **58**, 21-36 (2019). https://doi.org:10.1007/s00394-019-02095-1

11 Manach, C. *et al.* Addressing the inter-individual variation in response to consumption of plant food bioactives: Towards a better understanding of their role in healthy aging and cardiometabolic risk reduction. *Molecular Nutrition & Food Research* **61**, 1600557 (2017). https://doi.org:10.1002/mnfr.201600557

12 Rothwell, J. A. *et al.* New biomarkers of coffee consumption identified by the non-targeted metabolomic profiling of cohort study subjects. *PLOS ONE* **9**, e93474 (2014). https://doi.org:10.1371/journal.pone.0093474

13 Gibbons, H. *et al.* Demonstration of the utility of biomarkers for dietary intake assessment; proline betaine as an example. *Molecular Nutrition & Food Research* **61**, 1700037 (2017). https://doi.org:10.1002/mnfr.201700037

14 Vázquez-Manjarrez, N. *et al.* Discovery and validation of banana intake biomarkers using untargeted metabolomics in human intervention and cross-sectional studies. *Journal of Nutrition* **149**, 1701-1713 (2019). https://doi.org:10.1093/jn/nxz125

15 Li, C. *et al.* Development and validation of a metabolite score for red meat intake: an observational cohort study and randomized controlled dietary intervention. *American Journal of Clinical Nutrition* **116**, 511-522 (2022). https://doi.org:10.1093/ajcn/nqac094

16 Rothwell, J. A. *et al.* A metabolomic study of biomarkers of habitual coffee intake in four European countries. *Molecular Nutrition & Food Research* **63**, 1900659 (2019). https://doi.org:10.1002/mnfr.201900659

17 Vol. Prevention, Pesticides and Toxic Substances (7508C) (ed United States Environmental Protection Agency) (1999).

18 Sri Harsha, P. S. C. *et al.* Biomarkers of legume intake in human intervention and observational studies: a systematic review. *Genes & Nutrition* **13**, 25 (2018). https://doi.org:10.1186/s12263-018-0614-6

19 Perera, T. *et al.* Identification and monitoring of metabolite markers of dry bean consumption in parallel human and mouse studies. *Molecular Nutrition & Food Research* **59**, 795-806 (2015). https://doi.org:10.1002/mnfr.201400847

20 Roca-Saavedra, P. *et al.* Phytanic acid consumption and human health, risks, benefits and future trends: A review. *Food Chemistry* **221**, 237-247 (2017). https://doi.org:10.1016/j.foodchem.2016.10.074

21 Wang, Y., Hodge, R. A., Stevens, V. L., Hartman, T. J. & McCullough, M. L. Identification and reproducibility of plasma metabolomic biomarkers of habitual food intake in a US diet validation study. *Metabolites* **10**, 382 (2020). https://doi.org:10.3390/metabo10100382

22 Cuparencu, C. *et al.* Biomarkers of meat and seafood intake: an extensive literature review. *Genes & Nutrition* **14**, 35 (2019). https://doi.org:10.1186/s12263-019-0656-4

23 Cuparencu, C. *et al.* The anserine to carnosine ratio: an excellent discriminator between white and red meats consumed by free-living overweight participants of the PREVIEW study. *European Journal of Nutrition* **60**, 179-192 (2021). https://doi.org:10.1007/s00394-020-02230-3

24 Hanhineva, K. *et al.* Nontargeted metabolite profiling discriminates diet-specific biomarkers for consumption of whole grains, fatty fish, and bilberries in a randomized controlled trial. *The Journal of Nutrition* **145**, 7-17 (2015). https://doi.org:10.3945/jn.114.196840

25 Rafiq, T. *et al.* Nutritional metabolomics and the classification of dietary biomarker candidates: a critical review. *Advances in Nutrition* **12**, 2333-2357 (2021). https://doi.org:10.1093/advances/nmab054

26 McCullough, M. L. *et al.* Metabolomic markers of healthy dietary patterns in US postmenopausal women. *American Journal of Clinical Nutrition* **109**, 1439-1451 (2019). https://doi.org:https://doi.org/10.1093/ajcn/nqy385

27 Low, D. Y. *et al.* Diet-related metabolites associated with cognitive decline revealed by untargeted metabolomics in a prospective cohort. *Molecular Nutrition & Food Research* **63**, 1900177 (2019). https://doi.org:10.1002/mnfr.201900177

28 Cheah, I. K., Tang, R. M. Y., Yew, T. S. Z., Lim, K. H. C. & Halliwell, B. Administration of pure ergothioneine to healthy human subjects: uptake, metabolism, and effects on biomarkers of oxidative damage and inflammation. *Antioxidants & Redox Signaling* **26**, 193-206 (2017). https://doi.org:10.1089/ars.2016.6778

29 Weigand-Heller, A. J., Kris-Etherton, P. M. & Beelman, R. B. The bioavailability of ergothioneine from mushrooms (Agaricus bisporus) and the acute effects on antioxidant capacity and biomarkers of inflammation. *Preventive Medicine* **54**, S75-S78 (2012). https://doi.org:10.1016/j.ypmed.2011.12.028

30 Kubec, R., Svobodová, M. & Velíšek, J. Gas-chromatographic determination of S-methylcysteine sulfoxide in cruciferous vegetables. *European Food Research and Technology* **213**, 386-388 (2001). https://doi.org:10.1007/s002170100384

31 Pujos-Guillot, E. *et al.* Mass spectrometry-based metabolomics for the discovery of biomarkers of fruit and vegetable intake: citrus fruit as a case study. *J Proteome Res* **12**, 1645-1659 (2013). https://doi.org:10.1021/pr300997c

32 Lloyd, A. J., Beckmann, M., Favé, G., Mathers, J. C. & Draper, J. Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. *British Journal of Nutrition* **106**, 812-824 (2011). https://doi.org:10.1017/S0007114511001164

33 Playdon, M. C. *et al.* Comparing metabolite profiles of habitual diet in serum and urine. *American Journal of Clinical Nutrition* **104**, 776-789 (2016). https://doi.org:10.3945/ajcn.116.135301

34 Guertin, K. A. *et al.* Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. *American Journal of Clinical Nutrition* **100**, 208-217 (2014). https://doi.org:10.3945/ajcn.113.078758

35 Irwig, M. S., El-Sohemy, A., Baylin, A., Rifai, N. & Campos, H. Frequent intake of tropical fruits that are rich in beta-cryptoxanthin is associated with higher plasma beta-cryptoxanthin concentrations in Costa Rican adolescents. *J Nutr* **132**, 3161-3167 (2002). https://doi.org:10.1093/jn/131.10.3161

36 Burri, B., La Frano, M. R. & Zhu, C. Absorption, metabolism, and functions of β-cryptoxanthin. *Nutrition Reviews* **74**, 69-82 (2016). https://doi.org:10.1093/nutrit/nuv064

37 Hanhineva, K. *et al.* Qualitative Characterization of Benzoxazinoid Derivatives in Whole Grain Rye and Wheat by LC-MS Metabolite Profiling. *Journal of Agricultural and Food Chemistry* **59**, 921-927 (2011). https://doi.org:10.1021/jf103612u

38 Meng, L. *et al.* Occurrence, temporal variation (2010–2018), distribution, and source appointment of Per- and Polyfluoroalkyl Substances (PFAS) in mollusks from the Bohai Sea, China. *ACS ES&T Water* **2**, 195-205 (2022). https://doi.org:10.1021/acsestwater.1c00346

39 Pallister, T. *et al.* Characterizing blood metabolomics profiles associated with self-reported food intakes in female twins. *PLOS ONE* **11**, e0158568 (2016). https://doi.org:10.1371/journal.pone.0158568

40 Gorgani, L., Mohammadi, M., Najafpour, G. D. & Nikzad, M. Piperine—the bioactive compound of black pepper: from isolation to medicinal formulations.

Comprehensive Reviews in Food Science and Food Safety **16**, 124-140 (2017). https://doi.org:10.1111/1541-4337.12246

41 Praticò, G., Gao, Q., Manach, C. & Dragsted, L. O. Biomarkers of food intake for Allium vegetables. *Genes & Nutrition* **13**, 34 (2018). https://doi.org:10.1186/s12263-018-0624-4

42 Chietra, W. *Forms, analysis and stability of selenium and vitamin E in flour and Asian noodles* Doctor of Philosophy thesis, RMIT UNiversity, (2015).

43 Schmidt, C. *et al.* Lactoyl leucine and isoleucine are bioavailable alternatives for canonical amino acids in cell culture media. *Biotechnology and Bioengineering* **118**, 3395-3408 (2021). https://doi.org:10.1002/bit.27755

44 Ulaszewska, M. *et al.* Food intake biomarkers for apple, pear, and stone fruit. *Genes & Nutrition* **13**, 29 (2018). https://doi.org:10.1186/s12263-018-0620-8

45 Fung, T. T. *et al.* Diet-quality scores and plasma concentrations of markers of inflammation and endothelial dysfunction. *American Journal of Clinical Nutrition* **82**, 163-173 (2005). https://doi.org:10.1093/ajcn.82.1.163

46 Chiuve, S. E. *et al.* Alternative dietary indices both strongly predict risk of chronic disease. *The Journal of Nutrition* **142**, 1009-1018 (2012). https://doi.org:10.3945/jn.111.157222

47 Appel, L. J. *et al.* A clinical trial of the effects of dietary patterns on blood pressure. DASH Collaborative Research Group. *New England Journal of Medicine* **336**, 1117-1124 (1997). https://doi.org:10.1056/nejm199704173361601

48 Bromage, S. *et al.* Development and validation of a novel food-based global diet quality score (GDQS). *Journal of Nutrition* **151**, 75s-92s (2021). https://doi.org:10.1093/jn/nxab244

49 Satija, A. *et al.* Healthful and unhealthful plant-based diets and the risk of coronary heart disease in U.S. adults. *Journal of the American College of Cardiology* **70**, 411-422 (2017). https://doi.org:10.1016/j.jacc.2017.05.047

50 Sadhu, N. *et al.* Metabolome-wide association identifies ferredoxin-1 (FDX1) as a determinant of cholesterol metabolism and cardiovascular risk in Asian populations. *Nature Cardiovascular Research* **4**, 567-583 (2025). https://doi.org:10.1038/s44161-025-00638-w

51 Nieuwenhuijse, H. in *Encyclopedia of Dairy Sciences (Third Edition)* (eds Paul L H McSweeney & John P McNamara) 549-554 (Academic Press, 2022).

52 Pantzaris, T. P. & Mohd, J. A. Palm Kernel Oil. 1-3 (Palm Oil Research Institute of Malaysia).

53 Dos Santos, N. J. A. *et al.* Effect of dietary palm kernel oil on the quality, fatty acid profile, and sensorial attributes of young bull meat. *Foods* **11** (2022). https://doi.org:10.3390/foods11040609

54 Pan, A. *et al.* Red meat consumption and risk of type 2 diabetes: 3 cohorts of US adults and an updated meta-analysis. *American Journal of Clinical Nutrition* **94**, 1088-1096 (2011). https://doi.org:10.3945/ajcn.111.018978

55 Liu, G. *et al.* Cooking methods for red meats and risk of type 2 diabetes: a prospective study of U.S. women. *Diabetes Care* **40**, 1041-1049 (2017). https://doi.org:10.2337/dc17-0204

56 Field, B. C., Gordillo, R. & Scherer, P. E. The role of ceramides in diabetes and cardiovascular disease regulation of ceramides by adipokines. *Frontiers in Endocrinology* **11**, 569250 (2020). https://doi.org:10.3389/fendo.2020.569250

57 Gibbons, H. *et al.* Demonstration of the utility of biomarkers for dietary intake assessment; proline betaine as an example. *Molecular Nutrition & Food Research* **61** (2017). https://doi.org:10.1002/mnfr.201700037

58 Rebholz, C. M., Lichtenstein, A. H., Zheng, Z., Appel, L. J. & Coresh, J. Serum untargeted metabolomic profile of the Dietary Approaches to Stop Hypertension (DASH) dietary pattern. *American Journal of Clinical Nutrition* **108**, 243-255 (2018). https://doi.org:10.1093/ajcn/nqy099

59 Gadgil, M. D. *et al.* Metabolomic profile of the Healthy Eating Index-2015 in the multiethnic study of atherosclerosis. *The Journal of Nutrition* (2023). https://doi.org:10.1016/j.tjnut.2023.05.030

60 Hannun, Y. A. & Obeid, L. M. Sphingolipids and their metabolism in physiology and disease. *Nature Reviews Molecular Cell Biology* **19**, 175-191 (2018). https://doi.org:10.1038/nrm.2017.107

61 Wu, G. Amino acids: metabolism, functions, and nutrition. *Amino Acids* **37**, 1-17 (2009). https://doi.org:10.1007/s00726-009-0269-0

62 Brennan, L. & de Roos, B. Role of metabolomics in the delivery of precision nutrition. *Redox Biology* **65**, 102808 (2023). https://doi.org:10.1016/j.redox.2023.102808

63 Eichelmann, F. *et al.* Lipidome changes due to improved dietary fat quality inform cardiometabolic risk reduction and precision nutrition. *Nature Medicine* (2024). https://doi.org:10.1038/s41591-024-03124-1

64 Wang, X. *et al.* The Health for Life in Singapore (HELIOS) Study: delivering Precision Medicine research for Asian populations. *medRxiv*, 2024.2005.2014.24307259 (2024). https://doi.org:10.1101/2024.05.14.24307259

65 Association, A. D. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care* **44**, S15-S33 (2020). https://doi.org:10.2337/dc21-S002

66 Organization, W. H. HEARTS D: diagnosis and management of type 2 diabetes. (2020).

67 Unger, T. *et al.* 2020 International Society of Hypertension Global Hypertension Practice Guidelines. *Hypertension* **75**, 1334-1357 (2020). https://doi.org:doi:10.1161/HYPERTENSIONAHA.120.15026

68 Whitton, C. *et al.* Relative validity and reproducibility of a food frequency questionnaire for assessing dietary intakes in a multi-ethnic asian population using 24-h eietary recalls and biomarkers. *Nutrients* **9** (2017). https://doi.org:10.3390/nu9101059

69 Whitton, C., Rebello, S. A., Lee, J., Tai, E. S. & van Dam, R. M. A healthy asian a posteriori dietary pattern correlates with a priori dietary patterns and is associated with cardiovascular disease risk factors in a multiethnic asian population. *The Journal of Nutrition* **148**, 616-623 (2018). https://doi.org:10.1093/jn/nxy016

70 Fung, T. T. *et al.* Adherence to a DASH-Style Diet and Risk of Coronary Heart Disease and Stroke in Women. *Archives of Internal Medicine* **168**, 713-720 (2008). https://doi.org:10.1001/archinte.168.7.713

71 Fung, T. T. *et al.* Diet-quality scores and plasma concentrations of markers of inflammation and endothelial dysfunction2. *The American Journal of Clinical Nutrition* **82**, 163-173 (2005). https://doi.org:https://doi.org/10.1093/ajcn/82.1.163

72 Chiuve, S. E. *et al.* Alternative dietary indices both strongly predict risk of chronic disease. *Journal of Nutrition* **142**, 1009-1018 (2012). https://doi.org:10.3945/jn.111.157222

73 Satija, A. *et al.* Healthful and unhealthful plant-based diets and the risk of coronary heart disease in U.S. adults. *Journal of the American College of Cardiology* **70**, 411-422 (2017). https://doi.org:10.1016/j.jacc.2017.05.047

74 Ministry of health Singapore. *Disease Burden*, <https://www.moh.gov.sg/others/resources-and-statistics/disease-burden> (2023).

75 Ford, L. *et al.* Precision of a clinical metabolomics profiling platform for use in the identification of inborn errors of metabolism. *Journal of Applied Laboratory Medicine* **5**, 342-356 (2020). https://doi.org:10.1093/jalm/jfz026

76 Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. 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 (2009). https://doi.org:10.1021/ac901536h

77 Evans, A. M. *et al.* High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. *Metabolomics* **4** (2014). https://doi.org:10.4172/2153-0769.1000132

78 Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**, 211-221 (2007). https://doi.org:10.1007/s11306-007-0082-2

79 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301-320 (2005). https://doi.org:10.1111/j.1467-9868.2005.00503.x

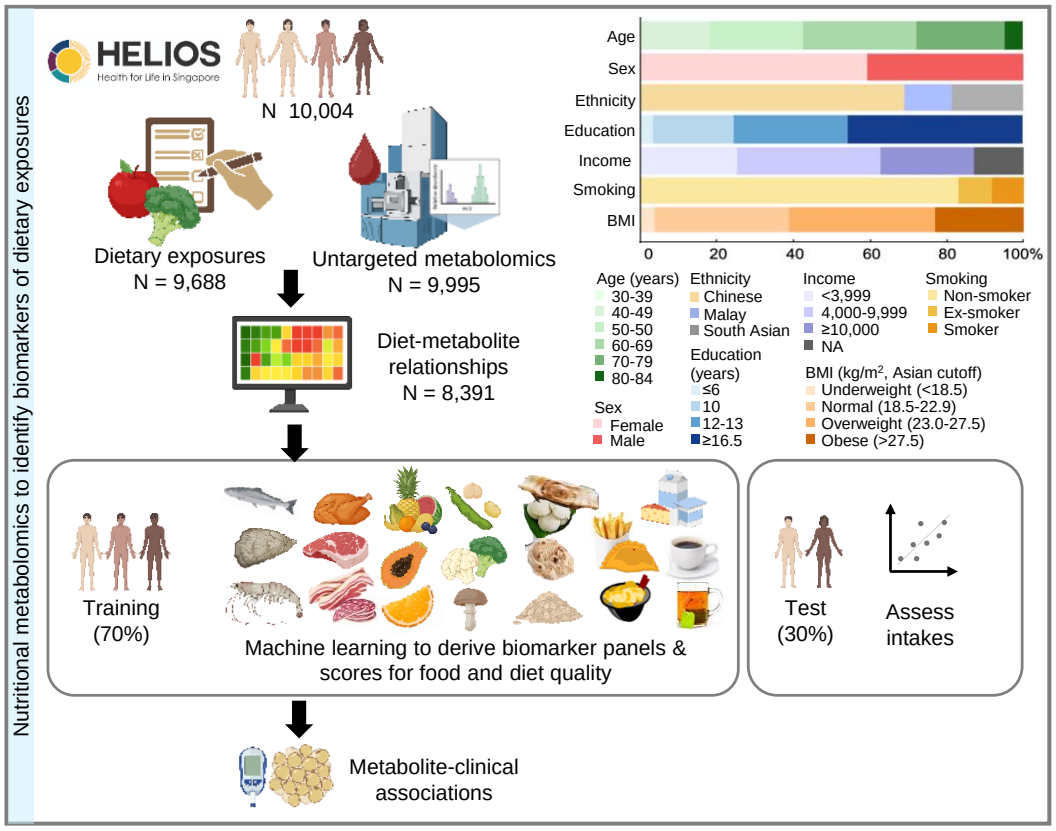
80 Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097-1103 (2021). https://doi.org:10.1038/s41588-021-00870-7

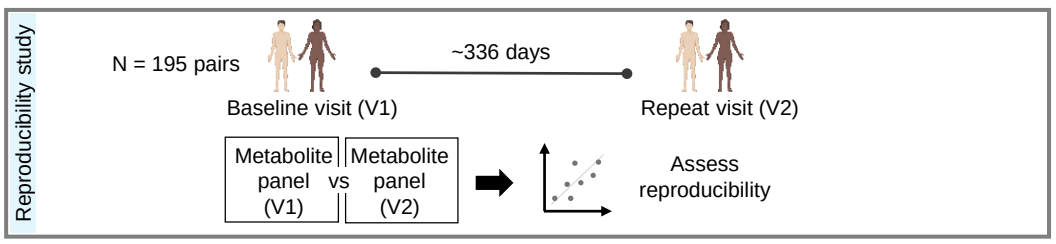
81 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191 (2010). https://doi.org:10.1093/bioinformatics/btq340

82 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015). https://doi.org:10.1186/s13742-015-0047-8

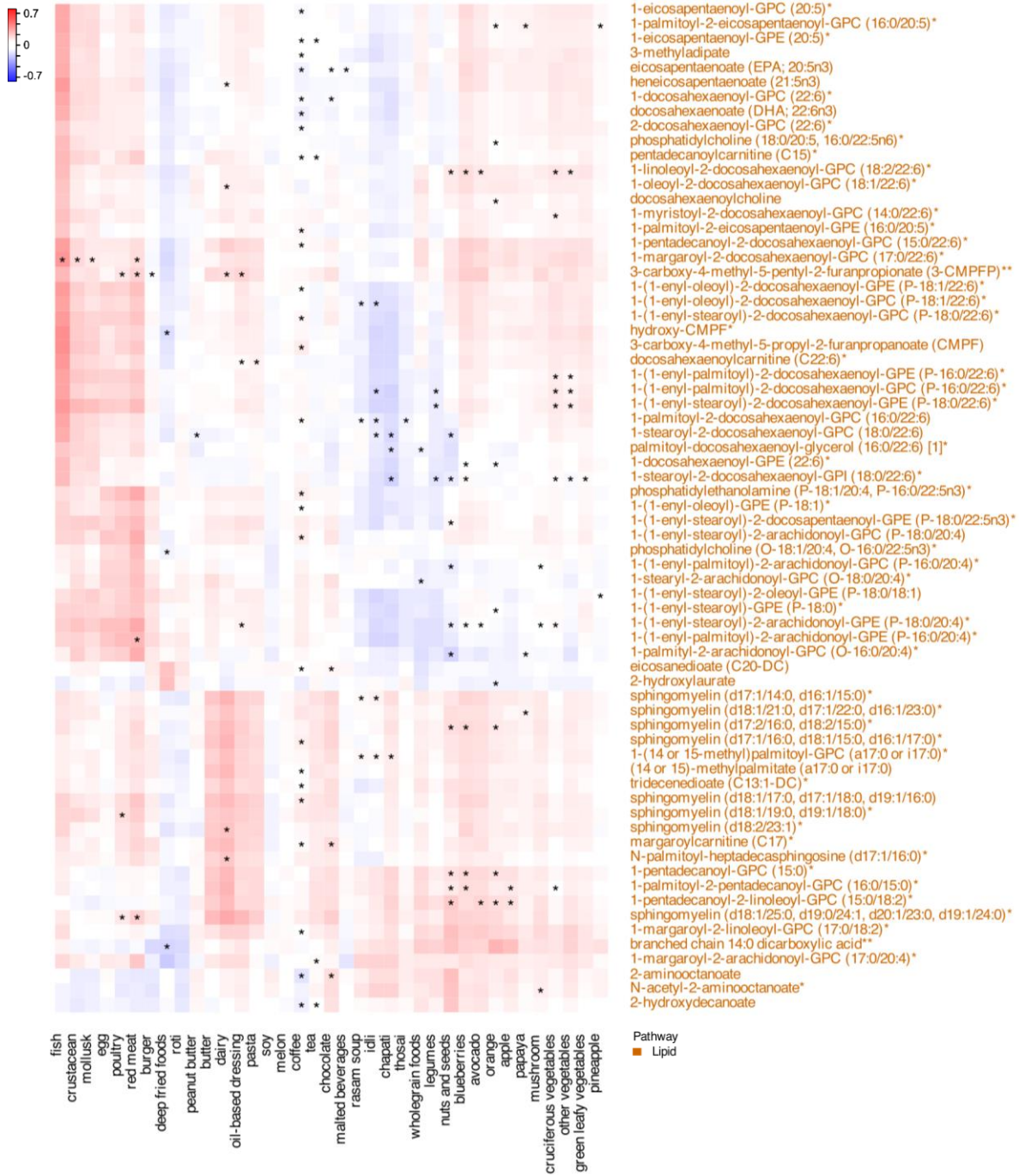
83 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* (2022). https://doi.org:10.1093/nar/gkac1010

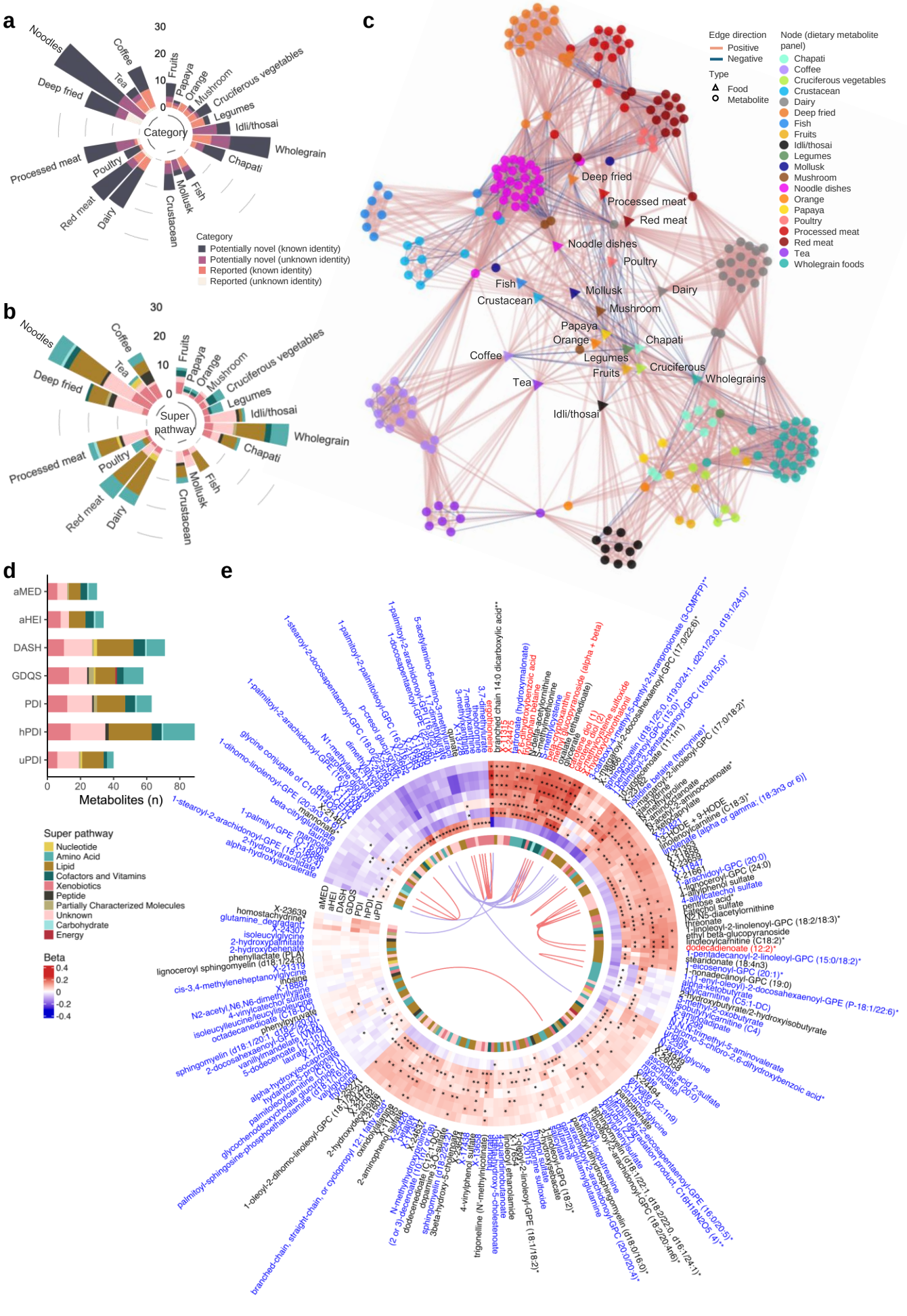
84 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607-d613 (2019). https://doi.org:10.1093/nar/gky1131

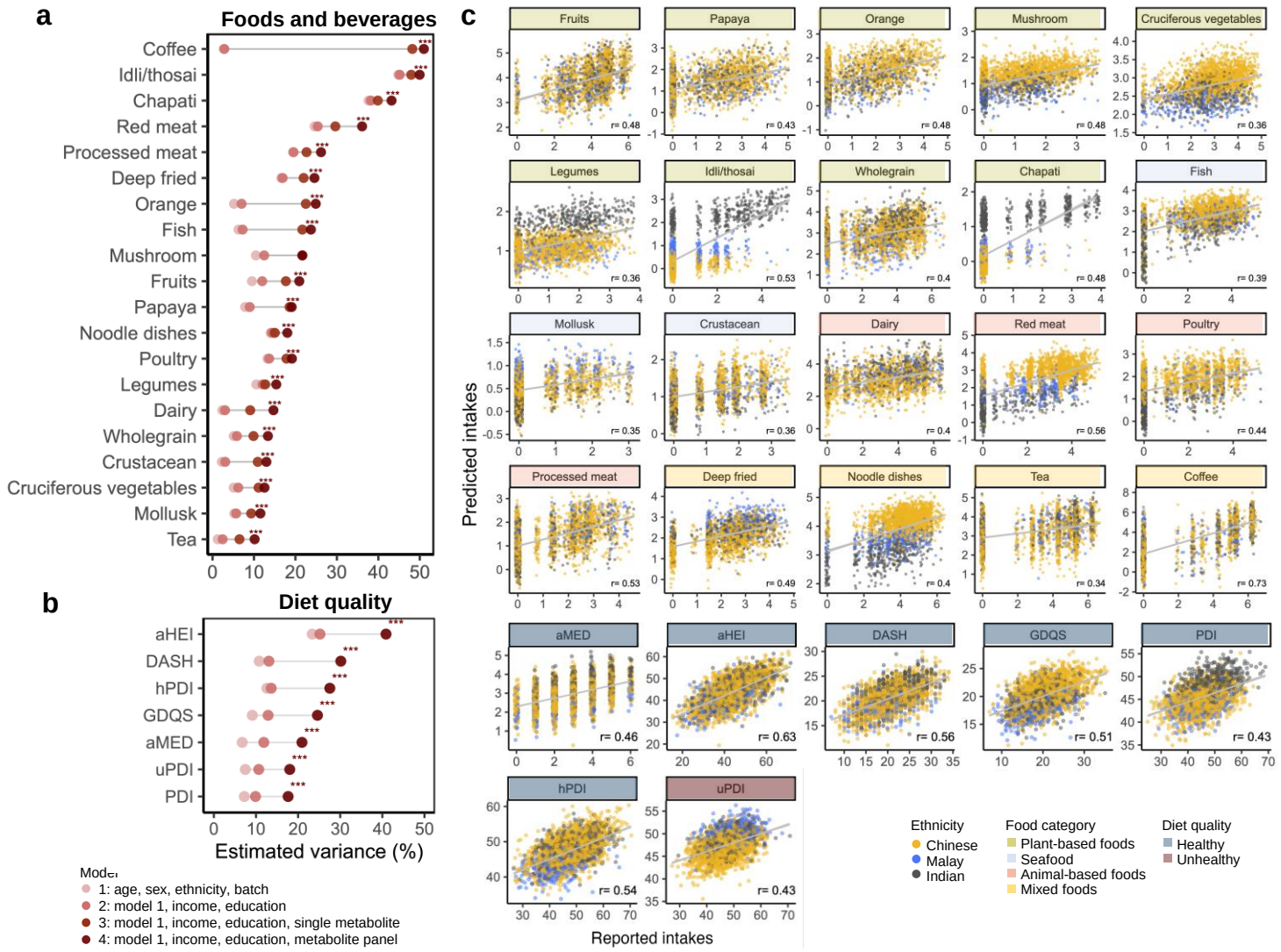

Nutritional metabolomics to identify biomarkers of dietary exposures
HELIOS
Health for Life in Singapore
N 10,004
Dietary exposures
N = 9,688
Untargeted metabolomics
N = 9,995
Diet-metabolite
relationships
N = 8,391
Age
Sex
Ethnicity
Education
Income
Smoking
BMI
0
20
40
60
80
100%
Age (years)
30-39
40-49
50-50
60-69
70-79
80-84
Ethnicity
Chinese
Malay
South Asian
Income
<3,999
4,000-9,999
≥10,000
NA
Smoking
Non-smoker
Ex-smoker
Smoker
Education
(years)
≤6
10
12-13
≥16.5
Sex
Female
Male
BMI (kg/m², Asian cutoff)
Underweight (<18.5)
Normal (18.5-22.9)
Overweight (23.0-27.5)
Obese (>27.5)
Training
(70%)
Machine learning to derive biomarker panels &
scores for food and diet quality
Test
(30%)
Assess
intakes
Metabolite-clinical
associations

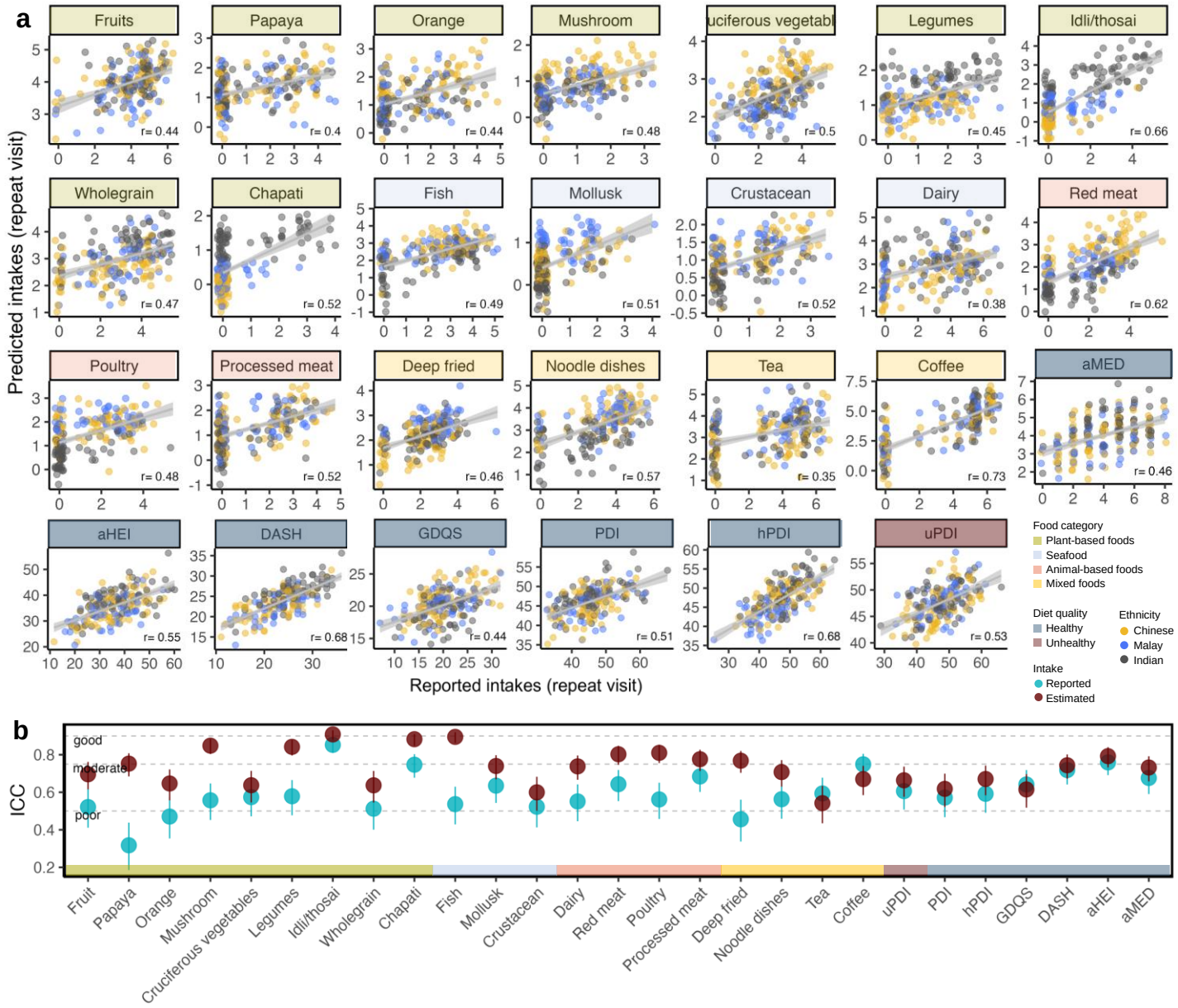

Reproducibility study
N = 195 pairs
~336 days
Baseline visit (V1)
Repeat visit (V2)
Metabolite
panel
(V1)
vs
Metabolite
panel
(V2)
Assess
reproducibility

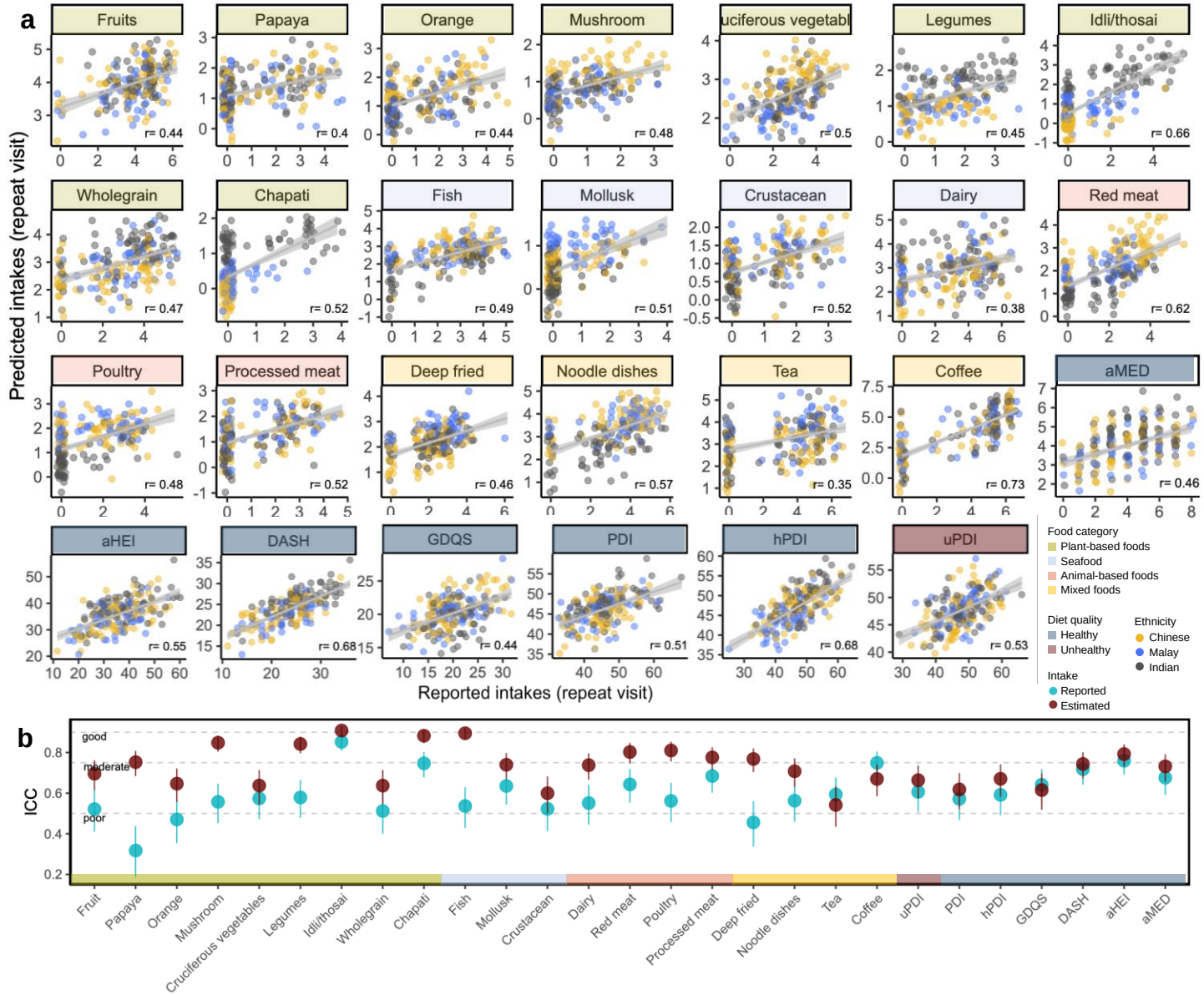

0.7
0
-0.7
oxalate (ethanedioate)
glycerate
tartronate (hydroxymalonate)
methyl glucopyranoside (alpha + beta)
X-11315
X-24475
carotene diol (1)
carotene diol (2)
N-delta-acetylornithine
N2,N5-diacetylornithine
4-hydroxychlorothalonil
S-methylcysteine sulfoxide
S-methylmethionine
S-methylcysteine
X-11847
X-21661
X-11858
X-11849
2,6-dihydroxybenzoic acid
tryptophan betaine
stachydrine
N-methylproline
N-methylhydroxyproline**
4-allylphenol sulfate
ethyl beta-glucopyranoside
X-25271
beta-cryptoxanthin
ergothioneine
histidine betaine (hercynine)*
X-11299
X-24556
phytanate
X-23974
X-13866
X-23782
2-ketocaprylate
X-11795
hippurate
2-aminophenol sulfate
X-23644
3-methoxycatechol sulfate (1)
X-17685
X-24637
3-bromo-5-chloro-2,6-dihydroxybenzoic acid*
3,5-dichloro-2,6-dihydroxybenzoic acid
N,N,N-trimethyl-5-aminovalerate
3,7-dimethylurate
X-13728
theobromine
7-methylxanthine
3-methylxanthine
4-ethylcatechol sulfate
4-vinylcatechol sulfate
o-cresol sulfate
pyroglutamylleucine*
X-25419
X-25810
X-25417
X-23587
perfluorooctanesulfonate (PFOS)
1-methyl-5-imidazolelactate
trans-4-hydroxyproline
4-hydroxyglutamate
X-11381
X-21383
X-16935
X-11308
X-21339
X-11372
X-11880
trigonelline (N'-methylnicotinate)
X-23639
homostachydrine*
theophylline
paraxanthine
1,7-dimethylurate
1-methylurate
1-methylxanthine
5-acetylamino-6-amino-3-methyluracil
1,3-dimethylurate
caffeine
1,3,7-trimethylurate
quinate
3-hydroxypyridine sulfate
cyclo(pro-val)
cyclo(leu-pro)
3-methyl catechol sulfate (1)
fish
crustacean
mollusk
egg
poultry
red meat
burger
deep fried foods
roti
peanut butter
butter
dairy
oil-based dressing
pasta
soy
melon
coffee
tea
chocolate
malted beverages
rasam soup
idli
chapati
thosai
wholegrain foods
legumes
nuts and seeds
blueberries
avocado
orange
apple
papaya
mushroom
other vegetables
cruciferous vegetables
green leafy vegetables
pineapple
Pathway
Amino acid
Carbohydrate
Cofactors and vitamins
Peptide
Unknown
Xenobiotics

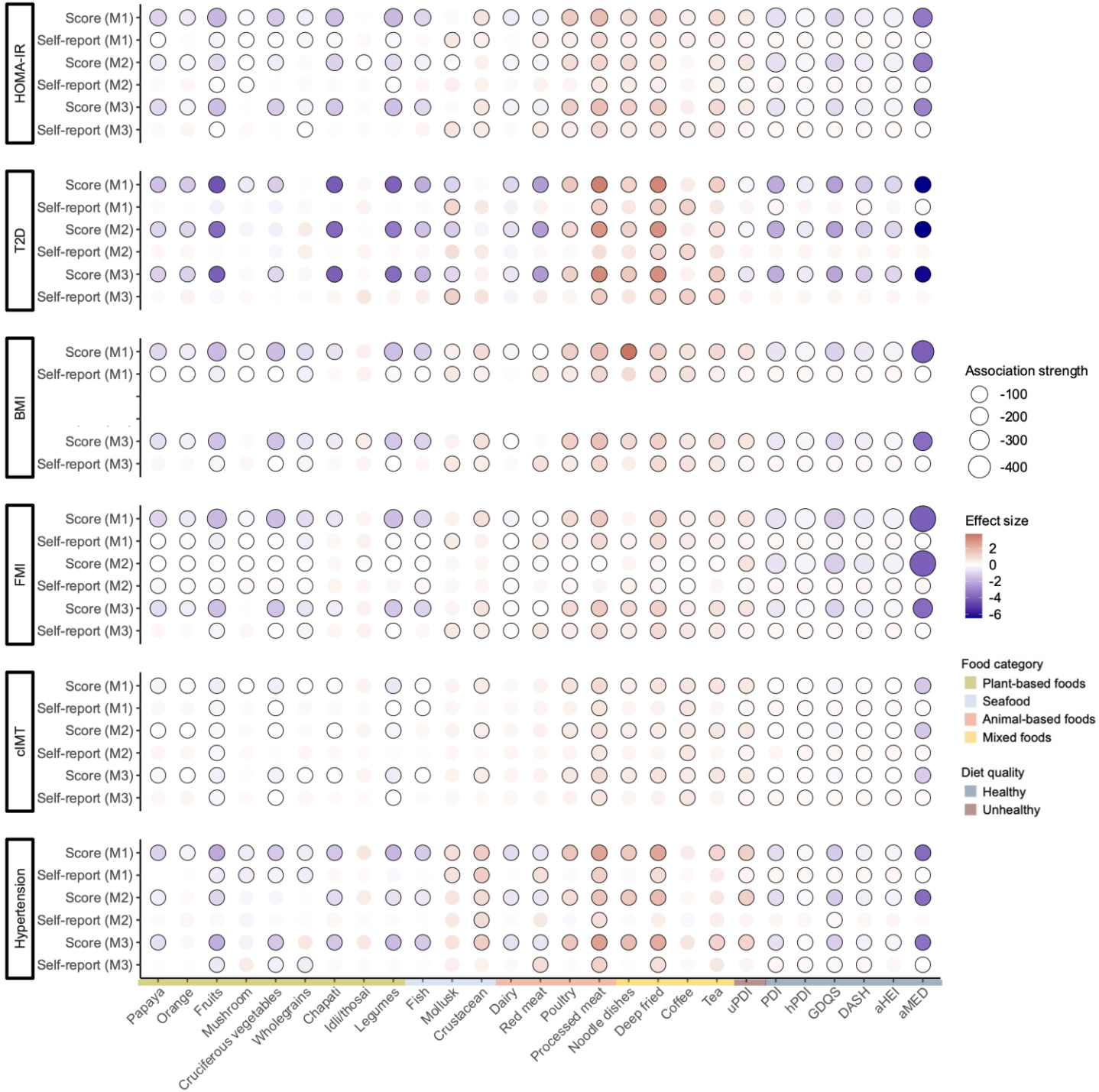

0.7
0
-0.7
1-eicosapentaenoyl-GPC (20:5)*
1-palmitoyl-2-eicosapentaenoyl-GPC (16:0/20:5)*
1-eicosapentaenoyl-GPE (20:5)*
3-methyladipate
eicosapentaenoate (EPA; 20:5n3)
heneicosapentaenoate (21:5n3)
1-docosahexaenoyl-GPC (22:6)*
docosahexaenoate (DHA; 22:6n3)
2-docosahexaenoyl-GPC (22:6)*
phosphatidylcholine (18:0/20:5, 16:0/22:5n6)*
pentadecanoylcarnitine (C15)*
1-linoleoyl-2-docosahexaenoyl-GPC (18:2/22:6)*
1-oleoyl-2-docosahexaenoyl-GPC (18:1/22:6)*
docosahexaenoylcholine
1-myristoyl-2-docosahexaenoyl-GPC (14:0/22:6)*
1-palmitoyl-2-eicosapentaenoyl-GPE (16:0/20:5)*
1-pentadecanoyl-2-docosahexaenoyl-GPC (15:0/22:6)*
1-margaroyl-2-docosahexaenoyl-GPC (17:0/22:6)*
3-carboxy-4-methyl-5-pentyl-2-furanpropionate (3-CMPFP)**
1-(1-enyl-oleoyl)-2-docosahexaenoyl-GPE (P-18:1/22:6)*
1-(1-enyl-oleoyl)-2-docosahexaenoyl-GPC (P-18:1/22:6)*
1-(1-enyl-stearoyl)-2-docosahexaenoyl-GPC (P-18:0/22:6)*
hydroxy-CMPF*
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)
docosahexaenoylcarnitine (C22:6)*
1-(1-enyl-palmitoyl)-2-docosahexaenoyl-GPE (P-16:0/22:6)*
1-(1-enyl-palmitoyl)-2-docosahexaenoyl-GPC (P-16:0/22:6)*
1-(1-enyl-stearoyl)-2-docosahexaenoyl-GPE (P-18:0/22:6)*
1-palmitoyl-2-docosahexaenoyl-GPC (16:0/22:6)
1-stearoyl-2-docosahexaenoyl-GPC (18:0/22:6)
palmitoyl-docosahexaenoyl-glycerol (16:0/22:6) [1]*
1-docosahexaenoyl-GPE (22:6)*
1-stearoyl-2-docosahexaenoyl-GPI (18:0/22:6)*
phosphatidylethanolamine (P-18:1/20:4, P-16:0/22:5n3)*
1-(1-enyl-oleoyl)-GPE (P-18:1)*
1-(1-enyl-stearoyl)-2-docosapentaenoyl-GPE (P-18:0/22:5n3)*
1-(1-enyl-stearoyl)-2-arachidonoyl-GPC (P-18:0/20:4)
phosphatidylcholine (O-18:1/20:4, O-16:0/22:5n3)*
1-(1-enyl-palmitoyl)-2-arachidonoyl-GPC (P-16:0/20:4)*
1-stearyl-2-arachidonoyl-GPC (O-18:0/20:4)*
1-(1-enyl-stearoyl)-2-oleoyl-GPE (P-18:0/18:1)
1-(1-enyl-stearoyl)-GPE (P-18:0)*
1-(1-enyl-stearoyl)-2-arachidonoyl-GPE (P-18:0/20:4)*
1-(1-enyl-palmitoyl)-2-arachidonoyl-GPE (P-16:0/20:4)*
1-palmityl-2-arachidonoyl-GPC (O-16:0/20:4)*
eicosanedioate (C20-DC)
2-hydroxylaurate
sphingomyelin (d17:1/14:0, d16:1/15:0)*
sphingomyelin (d18:1/21:0, d17:1/22:0, d16:1/23:0)*
sphingomyelin (d17:2/16:0, d18:2/15:0)*
sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0)*
1-(14 or 15-methyl)palmitoyl-GPC (a17:0 or i17:0)*
(14 or 15)-methylpalmitate (a17:0 or i17:0)
tridecenedioate (C13:1-DC)*
sphingomyelin (d18:1/17:0, d17:1/18:0, d19:1/16:0)
sphingomyelin (d18:1/19:0, d19:1/18:0)*
sphingomyelin (d18:2/23:1)*
margaroylcarnitine (C17)*
N-palmitoyl-heptadecasphingosine (d17:1/16:0)*
1-pentadecanoyl-GPC (15:0)*
1-palmitoyl-2-pentadecanoyl-GPC (16:0/15:0)*
1-pentadecanoyl-2-linoleoyl-GPC (15:0/18:2)*
sphingomyelin (d18:1/25:0, d19:0/24:1, d20:1/23:0, d19:1/24:0)*
1-margaroyl-2-linoleoyl-GPC (17:0/18:2)*
branched chain 14:0 dicarboxylic acid**
1-margaroyl-2-arachidonoyl-GPC (17:0/20:4)*
2-aminooctanoate
N-acetyl-2-aminooctanoate*
2-hydroxydecanoate
fish
crustacean
mollusk
egg
poultry
red meat
burger
deep fried foods
roti
peanut butter
butter
dairy
oil-based dressing
pasta
soy
melon
coffee
tea
chocolate
malted beverages
rasam soup
idli
chapati
thosai
wholegrain foods
legumes
nuts and seeds
blueberries
avocado
orange
apple
papaya
mushroom
cruciferous vegetables
other vegetables
green leafy vegetables
pineapple
Pathway
Lipid


a
Noodles
Coffee
Tea
Deep fried
Fruits
Papaya
Orange
Mushroom
Cruciferous vegetables
Legumes
Idli/thosai
Wholegrain
Chapati
Processed meat
Poultry
Red meat
Dairy
Fish
Mollusk
Crustacean
Category
Potentially novel (known identity)
Potentially novel (unknown identity)
Reported (known identity)
Reported (unknown identity)
b
Super pathway
c
Edge direction
Positive
Negative
Type
Food
Metabolite
Node (dietary metabolite panel)
Chapati
Coffee
Cruciferous vegetables
Crustacean
Dairy
Deep fried
Fish
Fruits
Idli/thosai
Legumes
Mollusk
Mushroom
Noodle dishes
Orange
Papaya
Poultry
Processed meat
Red meat
Tea
Wholegrain foods
Wholegrains
d
aMED
aHEI
DASH
GDQS
PDI
hPDI
uPDI
Metabolites (n)
Super pathway
Nucleotide
Amino Acid
Lipid
Cofactors and Vitamins
Xenobiotics
Peptide
Partially Characterized Molecules
Unknown
Carbohydrate
Energy
Beta
e

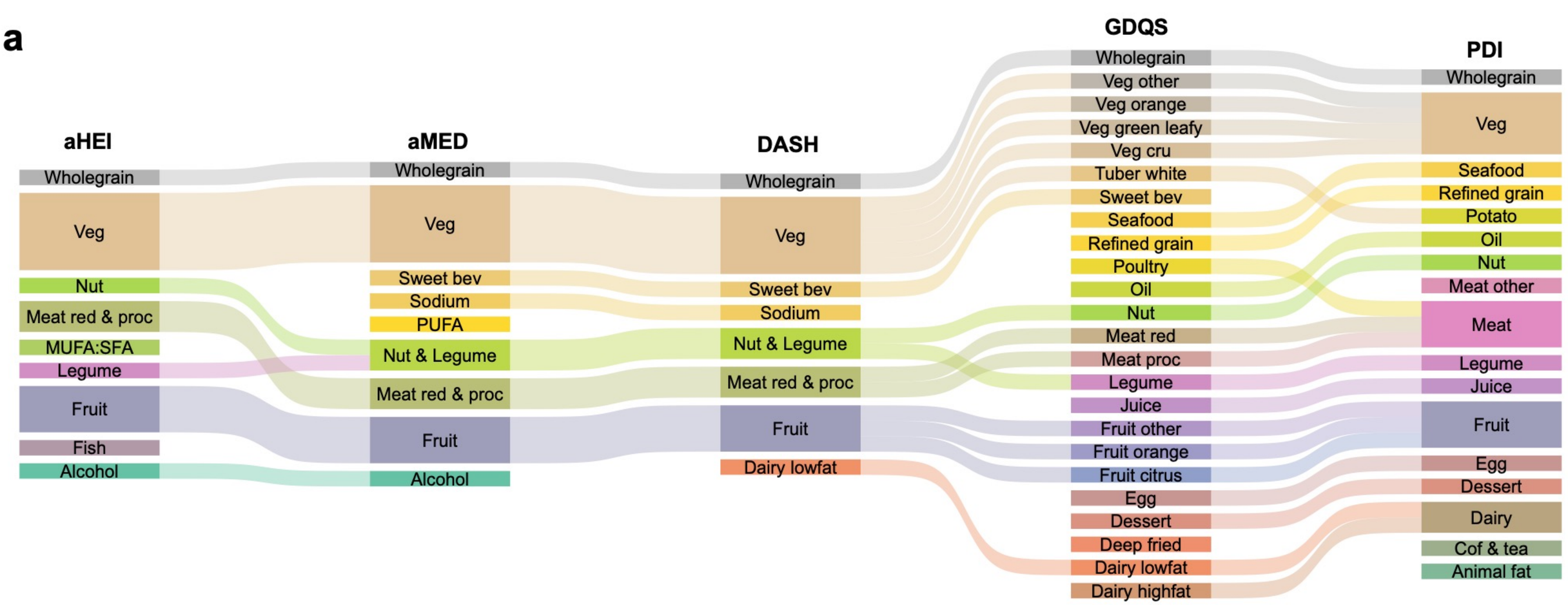

a
Foods and beverages
Coffee
Idli/thosai
Chapati
Red meat
Processed meat
Deep fried
Orange
Fish
Mushroom
Fruits
Papaya
Noodle dishes
Poultry
Legumes
Dairy
Wholegrain
Crustacean
Cruciferous vegetables
Mollusk
Tea
0 10 20 30 40 50
b
Diet quality
aHEI
DASH
hPDI
GDQS
aMED
uPDI
PDI
0 10 20 30 40 50
Estimated variance (%)
Model
1: age, sex, ethnicity, batch
2: model 1, income, education
3: model 1, income, education, single metabolite
4: model 1, income, education, metabolite panel
c
Fruits
Papaya
Orange
Mushroom
Cruciferous vegetables
Legumes
Idli/thosai
Wholegrain
Chapati
Fish
Mollusk
Crustacean
Dairy
Red meat
Poultry
Processed meat
Deep fried
Noodle dishes
Tea
Coffee
aMED
aHEI
DASH
GDQS
PDI
hPDI
uPDI
Predicted intakes
Reported intakes
Ethnicity
Chinese
Malay
Indian
Food category
Plant-based foods
Seafood
Animal-based foods
Mixed foods
Diet quality
Healthy
Unhealthy


a
Fruits
Papaya
Orange
Mushroom
uciferous vegetabl
Legumes
Idli/thosai
Wholegrain
Chapati
Fish
Mollusk
Crustacean
Dairy
Red meat
Poultry
Processed meat
Deep fried
Noodle dishes
Tea
Coffee
aMED
aHEI
DASH
GDQS
PDI
hPDI
uPDI
Predicted intakes (repeat visit)
Reported intakes (repeat visit)
Food category
Plant-based foods
Seafood
Animal-based foods
Mixed foods
Diet quality
Healthy
Unhealthy
Ethnicity
Chinese
Malay
Indian
Intake
Reported
Estimated
b
ICC
good
moderate
poor
Fruit
Papaya
Orange
Mushroom
Cruciferous vegetables
Legumes
Idli/thosai
Wholegrain
Chapati
Fish
Mollusk
Crustacean
Dairy
Red meat
Poultry
Processed meat
Deep fried
Noodle dishes
Tea
Coffee
uPDI
PDI
hPDI
GDQS
DASH
aHEI
aMED

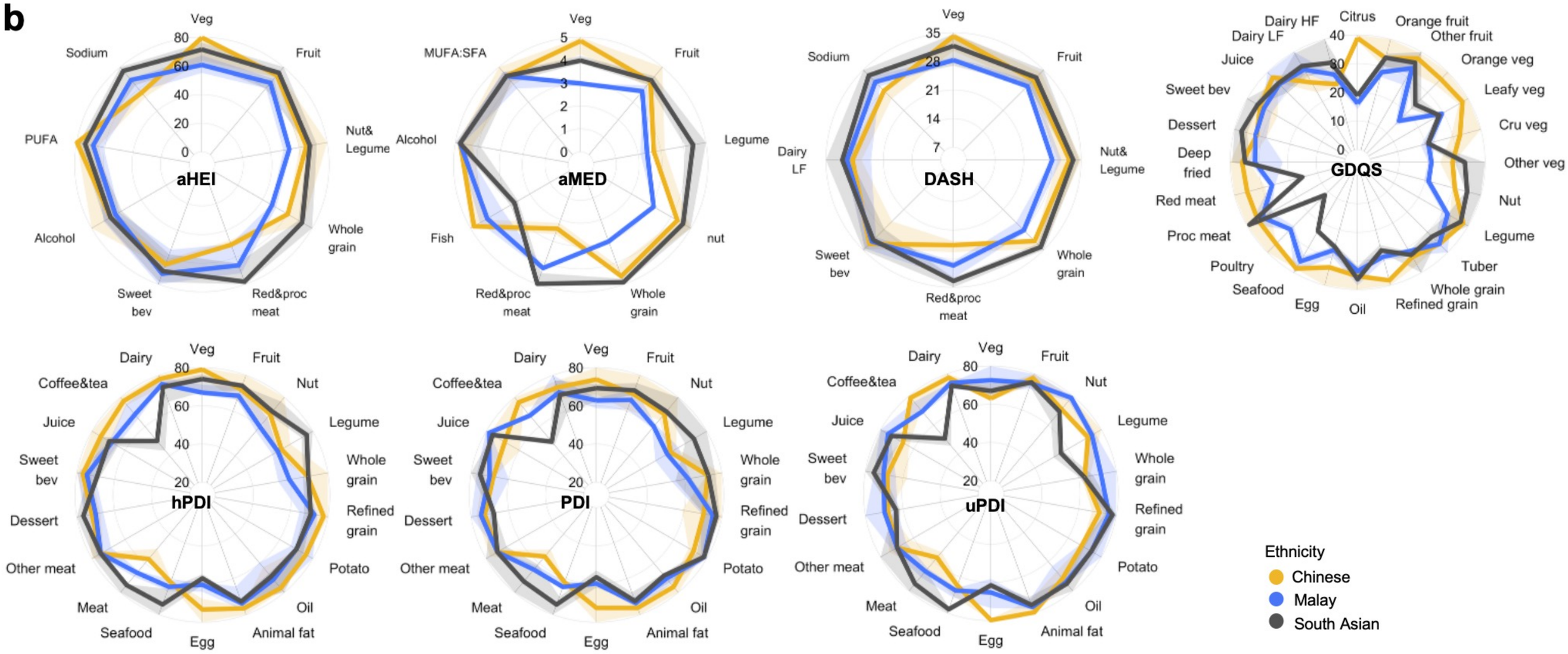

a
Fruits
Papaya
Orange
Mushroom
uciferous vegetabl
Legumes
Idli/thosai
Wholegrain
Chapati
Fish
Mollusk
Crustacean
Dairy
Red meat
Poultry
Processed meat
Deep fried
Noodle dishes
Tea
Coffee
aMED
aHEI
DASH
GDQS
PDI
hPDI
uPDI
r= 0.44
r= 0.4
r= 0.44
r= 0.48
r= 0.5
r= 0.45
r= 0.66
r= 0.47
r= 0.52
r= 0.49
r= 0.51
r= 0.52
r= 0.38
r= 0.62
r= 0.48
r= 0.52
r= 0.46
r= 0.57
r= 0.35
r= 0.73
r= 0.46
r= 0.55
r= 0.68
r= 0.44
r= 0.51
r= 0.68
r= 0.53
Predicted intakes (repeat visit)
Reported intakes (repeat visit)
Food category
Plant-based foods
Seafood
Animal-based foods
Mixed foods
Diet quality
Healthy
Unhealthy
Ethnicity
Chinese
Malay
Indian
Intake
Reported
Estimated
b
ICC
good
moderate
poor
Fruit
Papaya
Orange
Mushroom
Cruciferous vegetables
Legumes
Idli/thosai
Wholegrain
Chapati
Fish
Mollusk
Crustacean
Dairy
Red meat
Poultry
Processed meat
Deep fried
Noodle dishes
Tea
Coffee
uPDI
PDI
hPDI
GDQS
DASH
aHEI
aMED


HOMA-IR
T2D
BMI
FMI
cIMT
Hypertension
Score (M1)
Self-report (M1)
Score (M2)
Self-report (M2)
Score (M3)
Self-report (M3)
Papaya
Orange
Fruits
Mushroom
Cruciferous vegetables
Wholegrains
Chapati
Idli/dosai
Legumes
Fish
Mollusk
Crustacean
Dairy
Red meat
Poultry
Processed meat
Noodle dishes
Deep fried
Coffee
Tea
uPDI
PDI
hPDI
GDQS
DASH
aHEI
aMED
Association strength
-100
-200
-300
-400
Effect size
2
0
-2
-4
-6
Food category
Plant-based foods
Seafood
Animal-based foods
Mixed foods
Diet quality
Healthy
Unhealthy

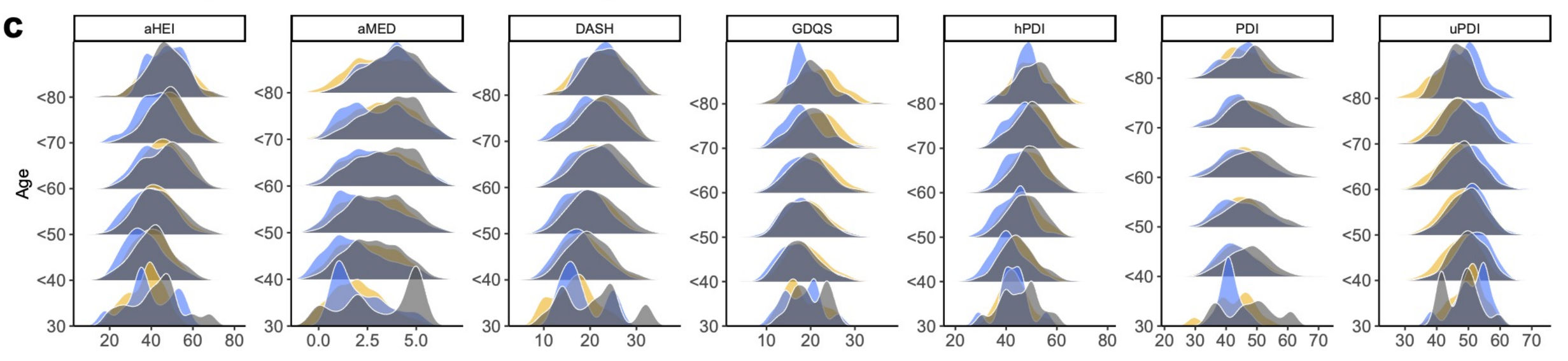
a


aHEI
Wholegrain
Veg
Nut
Meat red & proc
MUFA:SFA
Legume
Fruit
Fish
Alcohol
aMED
Wholegrain
Veg
Sweet bev
Sodium
PUFA
Nut & Legume
Meat red & proc
Fruit
Alcohol
DASH
Wholegrain
Veg
Sweet bev
Sodium
Nut & Legume
Meat red & proc
Fruit
Dairy lowfat
GDQS
Wholegrain
Veg other
Veg orange
Veg green leafy
Veg cru
Tuber white
Sweet bev
Seafood
Refined grain
Poultry
Oil
Nut
Meat red
Meat proc
Legume
Juice
Fruit other
Fruit orange
Fruit citrus
Egg
Dessert
Deep fried
Dairy lowfat
Dairy highfat
PDI
Wholegrain
Veg
Seafood
Refined grain
Potato
Oil
Nut
Meat other
Meat
Legume
Juice
Fruit
Egg
Dessert
Dairy
Cof & tea
Animal fat

b


aHEI
Veg
Fruit
Nut& Legume
Whole grain
Red&proc meat
Sweet bev
Alcohol
PUFA
Sodium
aMED
Veg
Fruit
Legume
nut
Whole grain
Red&proc meat
Fish
Alcohol
MUFA:SFA
DASH
Veg
Fruit
Nut& Legume
Whole grain
Red&proc meat
Sweet bev
Dairy LF
Sodium
GDQS
Veg
Orange veg
Leafy veg
Cru veg
Other veg
Nut
Legume
Tuber
Whole grain
Refined grain
Oil
Egg
Seafood
Poultry
Proc meat
Red meat
Deep fried
Dessert
Sweet bev
Juice
Dairy LF
Dairy HF
Citrus
Orange fruit
Other fruit
hPDI
Veg
Fruit
Nut
Legume
Whole grain
Refined grain
Potato
Oil
Animal fat
Egg
Seafood
Meat
Other meat
Dessert
Sweet bev
Juice
Coffee&tea
Dairy
PDI
Veg
Fruit
Nut
Legume
Whole grain
Refined grain
Potato
Oil
Animal fat
Egg
Seafood
Meat
Other meat
Dessert
Sweet bev
Juice
Coffee&tea
Dairy
uPDI
Veg
Fruit
Nut
Legume
Whole grain
Refined grain
Potato
Oil
Animal fat
Egg
Seafood
Meat
Other meat
Dessert
Sweet bev
Juice
Coffee&tea
Dairy
Ethnicity
Chinese
Malay
South Asian

c


aHEI
aMED
DASH
GDQS
hPDI
PDI
uPDI
Age
<80
<70
<60
<50
<40
30
20 40 60 80
0.0 2.5 5.0
10 20 30
10 20 30
20 40 60 80
20 30 40 50 60 70
30 40 50 60 70

a

Fruits
Papaya
Orange
Mushroom
Cruciferous vegetable
Legumes
Idli/thosai
Wholegrain
Chapati
Fish
Mollusk
Crustacean
Dairy
Red meat
Poultry
Processed meat
Deep fried
Noodle dishes
Tea
Coffee
aHEI
aMED
DASH
GDQS
PDI
hPDI
uPDI

Predictor error %

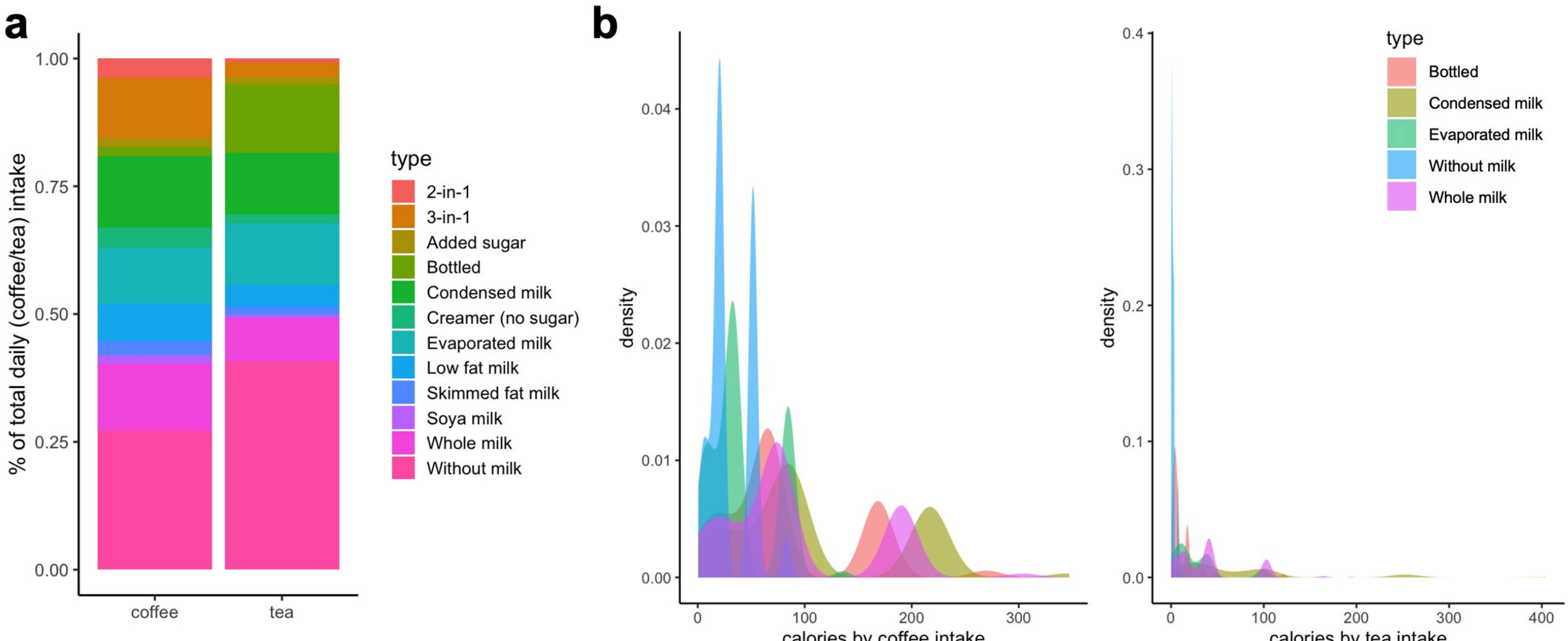
Reported intakes

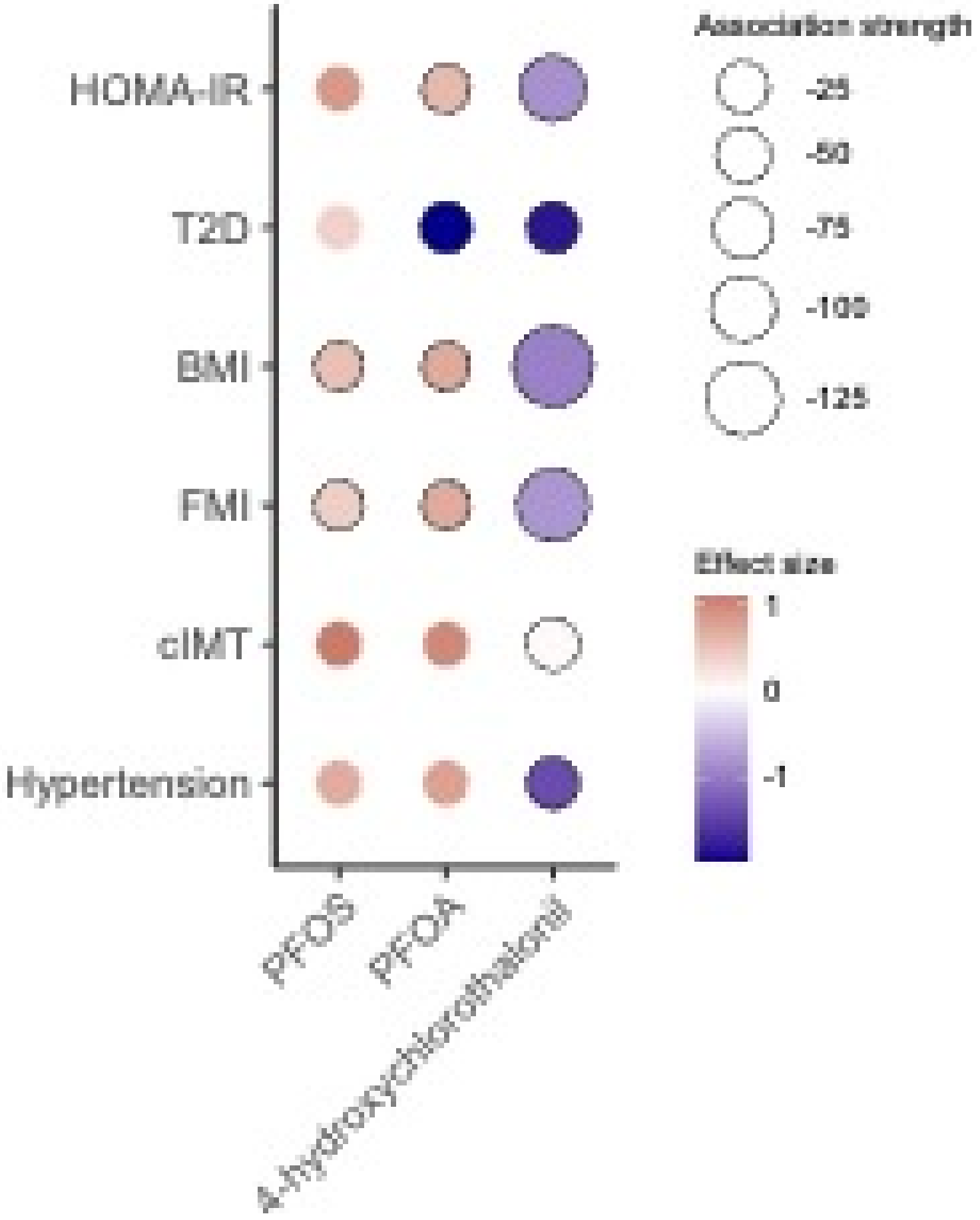
b

Fruits 83.8 %
Papaya 56.3 %
Orange 55.9 %
Mushroom 62.8 %
Cruciferous vegetable 78.8 %
Legumes 61.2 %
Idli/thosai 57.4 %
Wholegrain 73 %
Chapati 49.1 %
Fish 77.8 %
Mollusk 40 %
Crustacean 65.3 %
Dairy 68.5 %
Red meat 76 %
Poultry 63.8 %
Processed meat 63.7 %
Deep fried 76.1 %
Noodle dishes 84 %
Tea 72.5 %
Coffee 83.9 %
aHEI 87.9 %
aMED 70.2 %
DASH 86.2 %
GDQS 86.3 %
PDI 91.2 %
hPDI 91.2 %
uPDI 92.1 %

Density

Prediction error %


a
1.00
0.75
0.50
0.25
0.00
% of total daily (coffee/tea) intake
coffee
tea
type
2-in-1
3-in-1
Added sugar
Bottled
Condensed milk
Creamer (no sugar)
Evaporated milk
Low fat milk
Skimmed fat milk
Soya milk
Whole milk
Without milk
b
0.04
0.03
0.02
0.01
0.00
density
0
100
200
300
calories by coffee intake
0.4
0.3
0.2
0.1
0.0
density
0
100
200
300
400
calories by tea intake
type
Bottled
Condensed milk
Evaporated milk
Without milk
Whole milk


HOMA-IR
T2D
BMI
FMI
cIMT
Hypertension
PFOS
PFOA
4-Hydroxychlorothalonil
Association strength
-25
-50
-75
-100
-125
Effect size
1
0
-1

	Overall population (n=8,391, baseline visit)
Sociodemographic variables	
Age, years old	52.25 ± 11.54
Sex, n (%)	
- Female	4971 (59.2%)
- Male	3420 (40.8%)
Ethnicity, n (%)	
- Chinese	5783 (68.9%)
- Malay	1038 (12.4%)
- Indian	1570 (18.7%)
Education, total years	13.91 ± 3.35
Income category, n (%)	
- <S$4,000	2118 (25.3%)
- S$4,000-9,999	3127 (37.3%)
- ≥S$10,000	2036 (24.3%)
- not applicable (NA)	1085 (12.9%)
Smoking status, n (%)	
- Non-smoker	6938 (82.7%)
- Ex-smoker	729 (8.7%)
- Smoker	697 (8.3 %)
Biochemistry variables	
Fasting glucose, mM	5.08 ± 1.22
HbA1C, mmol/mol	5.71 ± 0.86
Total cholesterol, mM	5.22 ± 0.98
HDL-cholesterol, mM	1.53 ± 0.42
LDL-cholesterol, mM	3.12 ± 0.86
Triglycerides, mM	1.27 ± 0.82
Clinical phenotypes	
Homeostatic model assessment for insulin resistance (HOMA-IR), n	2.47 ± 2.23
Type 2 diabetes (T2D[1]), n (%)	1040 (12.4%)
- Medication use	509 (6.1%)
BMI, kg/m^2	24.78 ± 4.58
Fat mass index, kg/m^2	8.43 ± 3.53
Carotid intima-media thickness (cIMT), n	0.60 ± 0.13
Hypertension[2], n (%)	2411 (28.7%)
- Medication use	1,253 (14.9%)

[1]Individuals were classified with T2D based on ≥7 mM fasting glucose level, ≥6.5% HbA1C, prior diagnosis by doctor (self-reported history) or taking diabetes medications. [2]Individuals were classified with hypertension based on systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, prior diagnosis by doctor (self-reported history) or taking hypertensive medications.

	Subset population (n=195, baseline visit)	Subset population (n=195, repeat visit)
Sociodemographic variables		
Age, years old	51.33 ± 11.71	52.90 ± 11.67
Sex, n (%)		
- Female	103 (52.8%)	
- Male	92 (47.2%)	
Ethnicity, n (%)		
- Chinese	73 (37.4%)	
- Malay	50 (30.3%)	
- Indian	72 (36.9%)	
Education, total years	14.58 ± 3.15	
Income category, n (%)		
- <S$4,000	44 (22.6%)	
- S$4,000-9,999	77 (39.5%)	
- ≥S$10,000	54 (27.7%)	
- not applicable (NA)	20 (10.3%)	
Smoking status, n (%)		
- Non-smoker	156 (80%)	
- Ex-smoker	18 (9.2%)	
- Smoker	21 (10.8%)	
Biochemistry variables		
Fasting glucose, mM	5.00 ± 1.22	5.07 ± 1.10
HbA1C, mmol/mol	5.80 ± 0.79	5.86 ± 0.71
Total cholesterol, mM	5.17 ± 0.97	5.13 ± 0.89
HDL-cholesterol, mM	1.49 ± 0.42	1.47 ± 0.40
LDL-cholesterol, mM	3.12 ± 0.85	3.13 ± 0.80
Triglycerides, mM	1.22 ± 0.61	1.20 ± 0.69
Clinical phenotypes		
Homeostatic model assessment for insulin resistance (HOMA-IR), n	2.65 ± 2.33	2.83 ± 2.14
Type 2 diabetes (T2D[1]), n (%)	18 (9.2%)	30 (15.4%)
- Medication use	10 (5.1%)	11 (6.0%)
BMI, kg/m^2	25.52 ± 4.38	25.56 ± 4.45
Fat mass index, kg/m^2	8.75 ± 3.68	8.68 ± 3.57
Carotid intima-media thickness (cIMT), n	0.60 ± 0.13	0.60 ± 0.13
Hypertension[2], n (%)	57 (29.2%)	78 (40%)
- Medication use	26 (13.3%)	28 (15.2%)

[1]Individuals were classified with T2D based on ≥7 mM fasting glucose level, ≥6.5% HbA1C, prior diagnosis by doctor (self-reported history) or taking diabetes medications. [2]Individuals were classified with hypertension based on systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, prior diagnosis by doctor (self-reported history) or taking hypertensive medications.

Extended Data Table 3. Description of food and beverage groups in HELIOS population study.

Food and beverages		FFQ items
Seafood	Fish[1]	Raw fish, nigiri, dried fish, canned fish, white fish, oily fish
	Crustacean[1]	Prawn, crab
	Mollusk[1]	Squid, oyster
Animal	Poultry[1]	Fresh poultry
	Red meat[1]	Fresh pork (fat, lean), fresh beef (fat, lean), mutton (fresh, lean)
	Processed meat[1]	Poultry ham, processed poultry, cured pork
	Organs	Organs
	Dairy[1]	Butter, milk (non-fat, low-fat, whole), yoghurt (reduced fat, regular), cheese (reduced fat, regular)
Plant-based	Cruciferous vegetables[1]	Broccoli, cauliflower, chye sim, kailan, white cabbage, coleslaw
	Green leafy vegetables	Kang kong, spinach
	Starch vegetables	Lotus root, pumpkin, corn, yam, potato, begedil
	Other vegetables	Beansprouts, eggplant, gourd, celery, okra, carrot, capsicum, lettuce, cucumber, raw tomato, seaweed, fried onion, preserved vegetables
	Legumes[1]	Peas, long beans, other beans, dhal
	Mushroom[1]	Mushroom
	Fruits[1]	Orange[1], apple, avocado, banana, blueberries, dragonfruit, durian, grape, guava, kiwi, mango, melon, papaya[1], pear, pineapple, strawberries, dried fruit
	Soy-based foods	Tofu, soya milk (low sugar, regular), beancurd dessert
	Nuts and seeds	Nuts, seeds
	Oats	Oats
Mixed/ processed foods	Egg dishes	Egg (scrambled, steamed), egg dishes (with pork, chicken, seafood, beancurd, mixed vegetables)
	Noodle dishes[1]	Noodles in gravy, noodles in soup, dry noodles, fried Noodles
	Rice (plain)	Plain rice (white, unpolished, mixture)
	Rice (porridge)	Plain rice porridge (white, unpolished, mixture), flavoured rice porridge (white, unpolished, mixture)
	Rice (flavoured)	Flavoured rice, lontong, rice roll, idli[1], thosai[1]
	Rice (fried)	Fried rice, fried rice with vegetables
	Pasta	Pasta
	Bread (refined)	White bread
	Bread (wholemeal)	Wholemeal bread
	Bread (others)	Fruit bread, french toast, pizza, murtabak, sweet-filled buns, savoury-filled buns, roti, chapati[1]
	Wholegrain[1]	Wholemeal bread, wholemeal pasta, oats
	Cereals	Cereals (3in1), flavoured cereals, mixed cereals, plain cereals
	Spread	Margarine, butter, kaya, peanut butter, jam, hazelnut spread
	Deep fried foods[1]	French fries, fried chips, deep fried sweet snacks, puri
	Snacks and desserts	Biscuits (sweet, savoury), biscuits with filling, pastry, puffs and pies, kueh desserts with coconut, kueh desserts without coconut, desserts in soup with coconut, desserts in soup without coconut, pancake, cake, ice cream, dumpling, chocolate
	Fast food	Burger, french fries
	Coffee[1]	Coffee with milk (condensed, evaporated, whole, low-fat, skimmed, soya), creamer no sugar, 3in1, 2in1), no milk, bottled
	Tea[1]	Tea with milk (condensed, evaporated, whole, low-fat, skimmed, soya), creamer no sugar, 3in1, 2in1), no milk, bottled
	Malted beverages	Malted beverages with milk (condensed, evaporated, whole, low-fat, skimmed, soya), creamer no sugar, 3in1, 2in1, no milk
	Sweetened beverages	Fruit or vegetable juice (100% juice), fruit or vegetable juice (not 100% juice), other non-carbonated sweetened drinks, carbonated soft drinks (sugar free), carbonated soft drinks (regular), sports drinks, milkshake, yoghurt drinks, powdered nutrition drinks
	Alcohol	Alcoholic beverages (beer, wine, liqueur)
	Dressing	Oil-based dressing, cream-based dressing
	Condiments	Gravy with coconut, gravy without coconut
	Soup	Cream soup, rasam soup, other clear soup

[1]Food/beverage groups whose metabolite panels have reasonable explanatory power (adjusted R^2>0.1), adjusted for age, sex, ethnicity and batch.