Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations[1–3]

Kristin A Guertin, Steven C Moore, Joshua N Sampson, Wen-Yi Huang, Qian Xiao, Rachael Z Stolzenberg-Solomon, Rashmi Sinha, and Amanda J Cross

ABSTRACT

Background: Metabolomics is an emerging field with the potential to advance nutritional epidemiology; however, it has not yet been applied to large cohort studies.

Objectives: Our first aim was to identify metabolites that are biomarkers of usual dietary intake. Second, among serum metabolites correlated with diet, we evaluated metabolite reproducibility and required sample sizes to determine the potential for metabolomics in epidemiologic studies.

Design: Baseline serum from 502 participants in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial was analyzed by using ultra-high-performance liquid-phase chromatography with tandem mass spectrometry and gas chromatography–mass spectrometry. Usual intakes of 36 dietary groups were estimated by using a food-frequency questionnaire. Dietary biomarkers were identified by using partial Pearson's correlations with Bonferroni correction for multiple comparisons. Intraclass correlation coefficients (ICCs) between samples collected 1 y apart in a subset of 30 individuals were calculated to evaluate intraindividual metabolite variability.

Results: We detected 412 known metabolites. Citrus, green vegetables, red meat, shellfish, fish, peanuts, rice, butter, coffee, beer, liquor, total alcohol, and multivitamins were each correlated with at least one metabolite ($P < 1.093 \times 10^{-6}$; $r = -0.312$ to 0.398); in total, 39 dietary biomarkers were identified. Some correlations (citrus intake with stachydrine) replicated previous studies; others, such as peanuts and tryptophan betaine, were novel findings. Other strong associations included coffee (with trigonelline-*N*-methylnicotinate and quinate) and alcohol (with ethyl glucuronide). Intraindividual variability in metabolite levels (1-y ICCs) ranged from 0.27 to 0.89. Large, but attainable, sample sizes are required to detect associations between metabolites and disease in epidemiologic studies, further emphasizing the usefulness of metabolomics in nutritional epidemiology.

Conclusions: We identified dietary biomarkers by using metabolomics in an epidemiologic data set. Given the strength of the associations observed, we expect that some of these metabolites will be validated in future studies and later used as biomarkers in large cohorts to study diet-disease associations. The PLCO trial was registered at clinicaltrials.gov as NCT00002540. *Am J Clin Nutr* 2014;100:208–17.

INTRODUCTION

Diet is a modifiable risk factor for chronic disease; however, epidemiologic studies do not consistently support associations between specific foods or nutrients and disease endpoints. Most epidemiologic studies rely on self-reported dietary assessment methods that are subject to recall bias and measurement error (1–3). There is a pressing need for dietary biomarkers to better capture exposure; however, few have been identified to date (4).

Metabolomics, the measurement of small molecules in biofluids, may more precisely define dietary exposures and thus provide better estimates of disease risk in epidemiologic studies. Metabolomics accounts for variability in metabolism, because of lifestyle or genetics for example, by measuring downstream components or metabolic products of foods; therefore, metabolites may better reflect "true exposure." Metabolites may also capture exposure to nonnutritive substances, such as pesticides and compounds generated by cooking (5), which may play important roles in disease etiology.

Untargeted metabolomics in small dietary intervention and cohort studies has identified some novel potential dietary biomarkers (6). Although recent studies have shown that metabolomics can be successfully applied to dietary research (7), most studies (8–10) were small dietary interventions. Traditionally, dietary biomarkers have been identified and validated in feeding studies, but markers thus identified may not perform well as proxies for usual food intake—the exposure considered to be most etiologically relevant—in a population study. If the biomarker has a short half-life or if the food of interest is consumed only infrequently levels detected at the time of actual biospecimen collection may not proxy usual intake. A recent citrus feeding study, for example, identified >600 ions associated with acute citrus consumption; however, only 12 ions were associated with usual dietary citrus consumption in a free-living population (10). A recent metabolomics study determined that groups of

[1] From the Nutritional Epidemiology Branch (KAG, SCM, QX, RZS-S, RS, and AJC), the Biostatistics Branch (JNS), and the Occupational and Environmental Epidemiology Branch (W-YH), Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Department of Health and Human Services, Bethesda, MD.

[2] Supported by the Intramural Research Program of the National Cancer Institute, NIH.

[3] Address correspondence to KA Guertin, National Cancer Institute, NIH, 9609 Medical Center Drive, Room 6E326, MSC 9760, Bethesda, MD 20892. E-mail: kristin.guertin@nih.gov.

Received October 30, 2013. Accepted for publication March 27, 2014.

First published online April 16, 2014; doi: 10.3945/ajcn.113.078758.

serum metabolites are associated with patterns of dietary intake, although the authors only investigated 127 metabolites which were limited to acylcarnitines and choline-containing phospholipids (11). An agnostic approach that measures hundreds of metabolites has the benefit of identifying novel findings that may not have been previously considered.

With the use of biospecimens and data collected from participants in the Prostate, Lung, Colorectal, and Ovarian (PLCO)[4] Cancer Screening Trial, our objectives were as follows: *1*) to identify serum metabolites that are correlated with self-reported dietary intake and *2*) to determine whether metabolomics is a promising and feasible tool to identify associations in nutritional epidemiology by determining metabolite reproducibility and required sample sizes for epidemiologic studies.

SUBJECTS AND METHODS

Study population

The PLCO Cancer Screening Trial is a multicenter randomized screening trial that randomly assigned >150,000 US men and women between 1993 and 2001 to a screening or control arm (12). Eligibility requirements included age 55–74 y at baseline and no previous history of prostate, lung, colorectal, or ovarian cancer. Demographic and lifestyle characteristics were assessed at baseline by a self-administered questionnaire.

We used metabolomics data from a nested case-control study within PLCO, the details of which are briefly presented here. Within the screening arm of the trial, individuals with the following characteristics were excluded: self-reported history of cancer at baseline (except for basal cell skin cancer) (n = 4924); <6 mo of follow-up (n = 168 additional individuals); rare cancer during follow-up (n = additional 1074 individuals); self-reported Crohn disease, ulcerative colitis, familial polyposis, Gardner syndrome, or colorectal polyps (n = additional 6429 individuals); or no baseline serum sample (n = additional 2866 individuals). Among remaining participants, those who completed baseline dietary and risk factor questionnaires and consented for biospecimen use (n = 52,705) were eligible for metabolomics assays; of these, 255 incident colorectal cancer cases (diagnosed at least 6 mo after baseline) and 254 matched controls were incidence-density sampled and matched according to age, sex, race, randomization year, and season of blood draw. Controls were alive and cancer-free at time of cancer diagnosis for the matched cases. Seven participants were excluded from analyses because of incomplete (≥8 missing responses; n = 5) and/or inaccurate (extreme caloric consumption; n = 3) dietary data from a food-frequency questionnaire (FFQ), resulting in a final sample size of 502 individuals. The PLCO trial was approved by the institutional review boards of the US National Cancer Institute and the 10 screening centers, and all participants provided informed consent. All participants (n = 502) contributed baseline serum samples; in addition, serum collected 1 y after baseline was also measured in 30 controls to calculate within-individual variability.

[4]Abbreviations used: FDR, false discovery rate; FFQ, food-frequency questionnaire; HEI, Healthy Eating Index; ICC, intraclass correlation coefficient; PLCO, Prostate, Lung, Colorectal, and Ovarian.

Dietary assessment

In the screening arm of the trial, usual dietary intake was assessed at baseline by using the National Cancer Institute's self-administered and validated 137-item FFQ (http://www3.cancer.gov/prevention/plco), which captured information on typical frequency of intake during the past year (13). Food items of unspecified content (eg, lasagna) were excluded from analyses, because correlations for unknown foods were considered uninformative for these purposes. Data on single-nutrient dietary supplements were excluded given that we were primarily interested in food sources of metabolites; however, we considered multivitamin use (yes or no) in our analyses. Thus, in total, 111 items of interest comprising food, beverage, and multivitamin/supplement intakes were considered in this study; analyses focused on 36 dietary groups constructed by combining food items with similar properties. These 36 categories were based primarily on the USDA's MyPlate classification, but categories were further divided according to biological components of foods (14). Dietary intake from the FFQ was converted to grams per day by multiplying self-reported frequency of intake by portion size; portion sizes were assigned a gram amount on the basis of national dietary data from the USDA (1994–1996 Continuing Survey of Food Intakes by Individuals) for each sex (15). For items for which serving size was not queried on the FFQ (eg, fruits and vegetables), an average ("medium") portion size was assigned. Hereafter, dietary intake refers to grams per day unless otherwise specified.

Diet quality was assessed by using the Healthy Eating Index (HEI) 2010 (16), which contains 12 components that capture an individual's compliance to the key 2010 *Dietary Guidelines for Americans* (17); 9 HEI components focus on the adequacy of the diet (eg, higher intakes of fruit, vegetables, greens, whole grains, dairy, total protein, seafood and plant protein, and fatty acids) and 3 focus on moderation (eg, lower consumption of refined grains, sodium, and empty calories) (16). We calculated HEI scores for each individual on the basis of his or her self-reported diet and according to established methods (16); possible scores range from 0 to 100, with higher scores indicating higher diet quality.

Metabolite assessment

Serum metabolites (~<1000 Da) were assayed from baseline serum samples by Metabolon Inc, whose platform and procedures have been previously described (18, 19). Briefly, ultra-high-performance liquid chromatography–mass spectrometry and tandem mass spectrometry, in addition to gas chromatography coupled with mass spectrometry, were used to identify peaks. Mass spectral peaks, retention times, and m/z were determined by using a chemical reference library generated from 2500 standards, and these values were used to determine the identity of individual metabolites as well as their relative quantities.

One batch of 30 samples was analyzed each day, and batch and position within a batch were randomly assigned. Matched cases and controls were arranged as consecutive samples within a batch and the order of case compared with control was counterbalanced within each batch. Replicate aliquots from a separate source of pooled serum were randomly inserted into each batch at a level of 10% and served as blinded quality-control samples. In addition, a standard sample was inserted by Metabolon every sixth sample.

Although our case-control pairs were randomly assigned to their batch, metabolite values were batch normalized by dividing each individual metabolite value by the batch mean (of nonmissing values) to account for small day-to-day drifts in chromatogram performance. Metabolites were log-transformed (natural log), values below the detection threshold were set to the minimum observed value, and the distribution was centered; hereafter, metabolite level refers to these transformed values. Metabolites whose levels were below the detectable limit for >95% of study participants were excluded from analyses.

Statistical analysis

Demographic characteristics and dietary intakes for cases and controls were compared by using 2-sided statistical tests (chi-square test for categorical variables, Wilcoxon rank-sum test for continuous variables). Correlations between diet and metabolites were investigated through partial Pearson's correlations adjusted for age, sex, smoking status (current smokers compared with former/never smokers), and total energy intake (kcal/d). HEI correlations were additionally adjusted for recent (now or within the past 2 y) supplement use (multivitamin or single-nutrient supplement). To account for multiple comparisons, we used Bonferroni correction of the *P* values by the number of metabolites measured. Metabolite correlations with dietary groups were adjusted at 0.05/(number of known metabolites investigated × number of food items), and correlations with HEI score were adjusted at 0.05/number of known metabolites. Moreover, we estimated the total number of metabolites associated with each dietary group at given false discovery rates (FDRs) (20), given that Bonferroni correction could be overly conservative (21). Finally, we estimated the correlation between each food item and an optimized linear combination of metabolite levels. The coefficients for the linear combination were chosen by applying Lasso (22) regression to cross-validation training sets (penalty chosen by 10-fold cross-validation within the training set), and the correlation was based on comparing the resulting predicted values to the observed values in the cross-validation test sets. The final estimates, an average of the estimated correlations across all test sets, are therefore unbiased. The SE of the estimated correlation was calculated by using the "approximating ρ" method developed by Nadeau and Bengio (23).

We investigated the influence of self-reported frequency of intake and portion size on metabolite levels by creating box plots. Congruence between metabolite levels and dietary intake was assessed by Spearman rank correlations and κ statistics.

We conducted stratified analyses by sex, age, geographic location of study center, disease status, stage of colorectal cancer, and length of follow-up at time of colorectal cancer diagnosis. We used serum samples collected 1 y after baseline from 30 individuals to calculate intraclass correlation coefficients (ICCs), a measure of within-individual variability. The pooled quality-control serum samples were used to calculate intraassay CVs, which were averaged over all individuals to determine technical reproducibility by the laboratory. Analyses used SAS software, version 9.1.3 (SAS Institute), with the exception of the FDR plots, which used R (R Project, version 2.15.2) (24).

Variance components

We decomposed the total variance, σ_T^2, of each metabolite into 3 different components: the between-subject variance, σ_B^2, which was also considered the variance of the "usual" level in a population; the within-subject variance, σ_W^2, which reflected the 1-y variability around the "usual" level within an individual; the technical variance or laboratory reproducibility, σ_E^2, which was the expected variance from 2 identical samples: $\sigma_T^2 = \sigma_B^2 + \sigma_W^2 + \sigma_E^2$. We defined the ICC as follows:

$$ICC = (\sigma_B^2 + \sigma_W^2) \div \sigma_T^2 = 1 - (\sigma_E^2 \div \sigma_T^2) \qquad (1)$$

For each metabolite, variance components and corresponding ICCs were estimated from a mixed model.

Estimating sample size needed in future studies

Our second objective was to estimate the number of individuals needed for a 1:1 case:control study to have a power of 0.8 to detect an association between each metabolite and a disease, accounting for σ_W^2 and σ_E^2 and the testing of multiple metabolites. We focused on the metabolites that are most significantly associated with each dietary group. We defined the effect size for a given metabolite to be the RR of disease for an individual in the top quartile of the usual metabolite levels, as compared with the bottom quartile. We assumed that studies will use a *t* test, with the appropriate Bonferroni-corrected significance threshold, to test for an association between the disease and each metabolite. We then estimated the total number of individuals needed to detect a metabolite with a power of 0.8, given the within-individual variability and assumed effect size. Further details can be found elsewhere (25) and in the Supplemental Appendix under "Supplemental data" in the online issue.

RESULTS

Baseline demographic characteristics of the 502 participants are shown in **Table 1**. Mean age was 64 y, and the sample was largely white. The sample included primarily former (48%) or never (44%) smokers. Demographic characteristics were similar for cases and controls (Supplemental Table 1 under "Supplemental data" in the online issue). There were no differences in most characteristics; although there was a significant difference in BMI (P = 0.018), the magnitude of difference in mean (±SD) BMI (in kg/m^2; 28 ± 5 compared with 27 ± 4) was modest. Self-reported usual dietary intake, estimated from the FFQ, is shown in **Table 2**. There was no difference in self-reported dietary intake between cases and controls (Supplemental Table 2 under "Supplemental data" in the online issue). Given these similarities and because disease was not a primary interest of this study, we combined cases and controls in all subsequent analyses.

Identification of biomarkers of diet

We detected 412 metabolites of known identity and 231 metabolites of unknown identity (26–28). An additional 14 known metabolites were excluded from analyses because of nondetectable levels in >95% of individuals. Among the 643 metabolites analyzed, the median percentage of individuals with nondetectable levels was 4%. Correlations between all 36 dietary groups and known metabolites are shown in **Table 3**; all

TABLE 1
Demographic characteristics of participants in a metabolomics study nested within the PLCO Cancer Screening Trial[1]

Characteristic	Value
Sex [*n* (%)]	
Men	281 (56)
Women	221 (44)
Age (y)	64 ± 5[2]
Race [*n* (%)]	
White	474 (94)
Smoking status [*n* (%)]	
Current	42 (8)
Former	240 (48)
Never	220 (44)
Education[3] [*n* (%)]	
High school or less	167 (33)
Post–high school/some college	159 (32)
College/postgraduate	175 (35)
Total energy intake (kcal/d)	2068 ± 819
BMI[4] (kg/m^2)	27 ± 5
HEI 2010[5]	
Total	52.81 (29.75–87.49)[6]
Quintile	
1	43.96 (29.75–47.55)
2	49.63 (47.58–51.11)
3	52.81 (51.13–54.87)
4	57.30 (54.94–63.74)
5	72.00 (63.82–87.49)

[1] *n* = 502. HEI, Healthy Eating Index; PLCO, Prostate, Lung, Colorectal, and Ovarian.
[2] Mean ± SD (all such values).
[3] One participant was missing data on education.
[4] Four participants were missing data on BMI.
[5] Possible total HEI score: range of 0–110.
[6] Median; range in parentheses (all such values).

significant correlations are shown, as well as the strongest, albeit nonsignificant, correlations for dietary groups with no significant findings. We identified 39 correlations between dietary groups and known metabolites that were significant at the Bonferroni-corrected level of $P < 1.093 \times 10^{-6}$ [$P = 0.05/(111$ foods $\times$ 412 identified metabolites)]; these correlations represented 13 dietary groups including citrus, green vegetables, red meat, fish, shellfish, butter, peanuts, rice, coffee, beer, liquor, total alcohol, and multivitamins (Table 3). Most of the findings were for exogenous metabolites derived from their food sources.

Our strongest findings—those with a $P < 1 \times 10^{-10}$— were for multivitamins, citrus, fish, peanuts, coffee, and alcohol (Table 3). Multivitamins were correlated with serum vitamin E (positive correlation with α-tocopherol and a corresponding negative correlation with γ-tocopherol) and 2 vitamin B metabolites (pyridoxate and pantothenate). Citrus fruit were correlated with stachydrine ($r = 0.398$), chiro-inositol ($r = 0.301$), scyllo-inositol ($r = 0.298$), and *N*-methyl proline ($r = 0.298$). Fish was positively correlated with 3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid ($r = 0.322$) and moderately correlated with DHA ($r = 0.260$) and EPA ($r = 0.244$), 2 omega-3 fatty acids present in fish oils. In addition to correlations with fish, DHA was also correlated with rice consumption ($r = 0.270$) and 3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid was correlated with green vegetable ($r = 0.222$) and shellfish ($r = 0.238$) consumption. Even food groups typically consumed in low quantities had significant associations with metabolites: for example, peanuts were associated with 2 metabolites, tryptophan betaine ($r = 0.352$) and 4-vinylphenol sulfate ($r = 0.279$). For those metabolites associated with particular food groups, the median metabolite levels increased as the self-reported frequency of intake and serving size increased (Supplemental Figure 1 under "Supplemental data" in the online issue). There was reasonable congruence between quartile of metabolite levels and quartile of

TABLE 2
Self-reported usual dietary consumption in a nested study within the PLCO Cancer Screening Trial[1]

Category and dietary group	Value
Fruit (g/d)	
Citrus: oranges, orange juice, grapefruit	109 (35–218)[2]
Berries: strawberries	3 (1–4)
Apples, pears	29 (14–60)
Melon: watermelon, cantaloupe	3 (1–7)
Bananas	31 (9–78)
Other: plums, apricots, peaches, prunes, raisins, grapes, pineapples	38 (19–62)
Vegetables (g/d)	
Cruciferous: broccoli, cabbage, Brussels sprouts, cauliflower, turnip greens, mustard greens, collards, kale, swiss chard	28 (15–51)
Greens: lettuce, spinach, green peppers	32 (18–55)
Yellow/orange vegetables: carrots, tomatoes, sweet potatoes, beets	72 (47–110)
Starchy vegetables: white potatoes, corn, peas	74 (42–113)
Alliums (garlic, onions)	9 (4–15)
Other: celery, green beans, squash, cucumbers	49 (27–74)
Meat/fish (g/d)	
Red meat (includes processed)	61 (32–105)
Poultry: chicken	21 (12–43)
Fish (excluding shellfish)	17 (9–32)
Shellfish	1 (0–2)
Processed meat: cold cuts, hot dogs, bacon, sausage	7 (3–19)
Snack foods (g/d)	
Baked sweets	20 (10–38)
Chocolate	1 (0–4)
Candy (nonchocolate)	1 (0–3)
Chips	2 (1–8)
Other foods (g/d)	
Tofu	0 (0–1)
Beans	19 (11–33)
Eggs	7 (3–19)
Added fats: butter, salad dressing, vegetable oil spreads, margarine	9 (4–17)
Butter	0 (0–2)
Peanuts	3 (1–9)
Rice (white)	11 (2–20)
Beverages (g/d)	
Dairy: milk	277 (158–473)
Coffee	843 (205–899)
Sugar-sweetened beverages: soda, fruit punch	9 (3–145)
Beer	7 (0–69)
Wine	1 (0–12)
Liquor	1 (0–4)
Total alcohol	2 (0–15)
Reported multivitamin use [*n* (%)][3]	241 (48)

[1] *n* = 502. PLCO, Prostate, Lung, Colorectal, and Ovarian.
[2] Median; 25th–75th percentile range in parentheses (all such values).
[3] Seven participants were missing report of multivitamin use.

TABLE 3
Top metabolites associated with dietary groups in a nested case-control study within the PLCO Cancer Screening Trial[1]

Category and dietary group	Metabolite	Correlation (r)	P
Fruit			
Citrus: oranges, orange juice, grapefruit	Stachydrine	0.398	2.23×10^{-20}*
	Chiro-inositol	0.301	7.28×10^{-12}*
	Scyllo-inositol	0.298	1.13×10^{-11}*
	N-methyl proline	0.298	1.15×10^{-11}*
Berries: strawberries	1-Palmitoylglycero-phospho-inositol	−0.132	3.00×10^{-3}
Apples, pears	13-HODE + 9-HODE	−0.141	2.00×10^{-3}
Melon: watermelon, cantaloupe	Pregnenolone sulfate	−0.156	4.82×10^{-4}
Bananas	γ-Tocopherol	−0.200	6.93×10^{-6}
Other: plums, apricots, peaches, prunes, raisins, grapes, pineapple	Pyridoxate	0.205	3.98×10^{-6}
Vegetables			
Cruciferous: broccoli, cabbage, Brussels sprouts, cauliflower, turnip greens, mustard greens, collards, kale, swiss chard	α-CEHC glucuronide	0.149	8.32×10^{-4}
Greens: lettuce, spinach, green peppers	CMPF	0.222	5.52×10^{-7}*
Yellow/orange vegetables: carrots, tomatoes, sweet potatoes, beets	Creatinine	0.123	6.00×10^{-3}
Starchy vegetables: white potatoes, corn, peas	Cyclo (-Leu-Pro)	−0.143	1.34×10^{-3}
Alliums (garlic, onions)	CMPF	0.182	4.23×10^{-5}
Other: celery, green beans, squash, cucumbers	DHA	0.161	3.22×10^{-4}
Meat/fish			
Red meat	Indolepropionate	−0.221	6.14×10^{-7}*
Poultry: chicken	Pyroglutamine	−0.176	7.77×10^{-5}
Fish (excluding shellfish)	CMPF	0.322	1.80×10^{-13}*
	DHA	0.260	3.87×10^{-9}*
	EPA	0.244	3.44×10^{-8}*
	1-Docosahexaenoylglycero-phosphocholine	0.237	8.27×10^{-8}*
Shellfish	CMPF	0.238	7.69×10^{-8}*
Processed meat: cold cuts, hot dogs, bacon, sausage	Lathosterol	0.180	5.39×10^{-5}
Snack foods			
Baked sweets	Glutamine	0.182	4.40×10^{-5}
Chocolate	Theobromine	0.164	2.28×10^{-4}
Candy (nonchocolate)	Leucylleucine	0.161	3.00×10^{-4}
Chips	DHA	−0.133	2.90×10^{-3}
Other			
Tofu	4-Ethylphenylsulfate	0.188	2.43×10^{-5}
Beans	*S*-Methylcysteine	0.168	1.72×10^{-4}
Eggs	Indolepropionate	−0.161	3.11×10^{-4}
Added fats: butter, salad dressing, vegetable oil spreads, margarine	δ-Tocopherol	0.192	1.55×10^{-5}
Butter[2]	Methyl palmitate (15 or 2)	0.262	2.97×10^{-9}*
	Pentadecanoate (15:0)	0.248	2.06×10^{-8}*
	10-Undecenoate (11:1n–1)	0.230	2.05×10^{-7}*
Peanuts	Tryptophan betaine	0.352	6.21×10^{-16}*
	4-Vinylphenol sulfate	0.279	2.39×10^{-10}*
Rice (white)	DHA	0.270	9.51×10^{-10}*
Beverages			
Dairy: milk	Homostachydrine	0.173	1.00×10^{-4}
Coffee	Trigonelline (N'-methylnicotinate)	0.424	3.36×10^{-23}*
	Quinate	0.372	8.00×10^{-18}*
	1-Methylxanthine	0.299	9.04×10^{-12}*
	Paraxanthine	0.270	8.87×10^{-10}*
	N-2-furoyl-glycine	0.264	2.30×10^{-9}*
	Catechol sulfate	0.232	1.58×10^{-7}*
Sugar-sweetened beverages: soda, fruit punch	Quinate	−0.177	7.21×10^{-5}
Beer	16-Hydroxypalmitate	0.221	6.30×10^{-7}*
Wine	Scyllo-inositol	0.200	7.19×10^{-6}
Liquor	Ethyl glucuronide	0.295	1.85×10^{-11}*
Total alcohol[3]	Ethyl glucuronide	0.360	1.04×10^{-16}*
	4-Androsten-3β,17β-diol disulfate 1	0.289	5.31×10^{-11}*

(Continued)

TABLE 3 (*Continued*)

Category and dietary group	Metabolite	Correlation (r)	P
	5-α-Androstan-3β,17β-diol disulfate	0.254	9.21×10^{-9}*
	Cyclo (-Leu-Pro)	0.249	1.84×10^{-8}*
	Bilirubin (E,Z or Z,E)	0.243	3.75×10^{-8}*
	16-Hydroxypalmitate	0.239	6.57×10^{-8}*
	Dihomo-linoleate (20:2n–6)	0.230	2.12×10^{-7}*
	Palmitoleate (16:1n–7)	0.230	2.15×10^{-7}*
Vitamins/supplements			
Multivitamins	Pantothenate	0.541	3.36×10^{-39}*
	Pyridoxate	0.433	3.67×10^{-24}*
	α-Tocopherol	0.368	2.04×10^{-17}*
	γ-Tocopherol	−0.312	1.02×10^{-12}*
	Threonate	0.268	1.16×10^{-9}*
	β-Tocopherol	−0.233	1.42×10^{-7}*

[1] $n = 502$. The untargeted approach investigated all identified metabolites and dietary groups captured by food-frequency questionnaire; all significant correlations are indicated with an asterisk. For dietary groups with no significant associations, only the strongest association is shown. Partial Pearson correlations adjusted for age, sex, smoking status (current smokers, former/never smokers), and total energy intake (kcal/d). Significance was defined as the Bonferroni-corrected level of $P < 1.093 \times 10^{-6}$ (111 food items × 412 identified metabolites, at the 0.05 level). CMPF, 3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid; PLCO, Prostate, Lung, Colorectal, and Ovarian; 9-HODE, 9-hydroxy-10,12-octadecadienoic acid; 13-HODE, 13-hydroxy-9,11-octadecadienoic acid; α-CEHC, 2,5,7,8-tetramethyl-2(2′-carboxyethyl)-6-hydroxychroman.

[2] Butter was also included in the added fats/oils group.

[3] Total alcohol is the combination of beer, wine, and liquor (in g/d); beer, wine, and liquor are also presented as separate groups.

dietary intake for the main associations we observed (Supplemental Table 3 under "Supplemental data" in the online issue).

Many beverages, including coffee, beer, liquor, and total alcohol, each had at least one significant metabolite biomarker, with much stronger correlations among liquor and total alcohol compared with beer (Table 3). Interestingly, although the metabolites associated with beer and liquor (16-hydroxypalmitate and ethyl glucuronide, respectively) were also associated with total alcohol intake, total alcohol (combining beer, wine, and liquor) had many significant correlations, some of which are involved in lipid pathways, which were not significant for the individual components. Coffee had very strong correlations with trigonelline N'-methylnicotinate ($r = 0.424$), quinate ($r = 0.372$), 1-methylxanthine ($r = 0.299$), and paraxanthine ($r = 0.270$), with an additional 2 moderate correlations with N-2-furoyl-glycine ($r = 0.264$) and catechol sulfate ($r = 0.232$).

We attempted to create better predictors of the questionnaire-based measures by combining information across multiple metabolites. For most food items, the correlation between the food item and the multimetabolite prediction score was of similar magnitude to the correlation between that food item and the most strongly associated metabolite (Supplemental Table 4 under "Supplemental data" in the online issue). However, there was a noted improvement when using the predictive score for butter [$r = 0.34$ for the multimetabolite score; $\max(r) = 0.26$ for a single metabolite], beer (0.31 compared with 0.22), and red meat (0.37 compared with 0.22).

We also analyzed the metabolites that we were unable to name. Twenty-four unidentified metabolites were significantly correlated with dietary groups (Supplemental Table 5 under "Supplemental data" in the online issue; with correlations ranging from −0.23 to 0.39); all food groups correlated with unknown metabolites were also significantly correlated with identified metabolites. In more detailed analyses, we also examined correlations between metabolites and the 111 individual food items that formed the dietary groups in the main analyses; results did not appreciably differ from the results for food groups (data not shown). In stratified analyses, the diet-metabolite correlations were consistent across sex (Supplemental Table 6 under "Supplemental data" in the online issue), age (Supplemental Table 6 under "Supplemental data" in the online issue), and geographical location of the study center (Supplemental Table 7 under "Supplemental data" in the online issue).

In sensitivity analyses in colorectal cancer cases and controls separately, the main correlations we observed between metabolites and dietary groups did not differ by disease status (Supplemental Table 8 under "Supplemental data" in the online issue). Overall, the direction and magnitude of diet-metabolite correlations did not differ appreciably by stage of colorectal cancer (Supplemental Table 9 under "Supplemental data" in the online issue) or length of follow-up time at colorectal cancer diagnosis (Supplemental Table 9 under "Supplemental data" in the online issue). In particular, associations between metabolites and citrus, coffee, and alcohol were robust in all sensitivity analyses.

Investigating dietary exposures by an overall index, we found that the median total HEI score was 53 (range: 30–87, out of possible 100 with a higher score indicating better diet quality) (Table 1). There were 5 significant metabolite correlations at the Bonferroni-corrected level of $P < 1.214 \times 10^{-4}$ ($P = 0.05/412$ identified metabolites) (**Table 4**), including a negative correlation with γ-tocopherol. Overall, correlations unadjusted for supplement use were similar to adjusted results but generally stronger, and 15 of the top 20 metabolite correlations remained among the strongest correlations regardless of adjustment (data not shown). Many of the correlations associated with HEI score represented vitamins, including vitamin E and constituents or metabolites of vitamins B (pantothenate, pyridoxate) and C (threonate).

TABLE 4
Top 20 metabolites associated with the HEI 2010 in a nested study within the PLCO Cancer Screening Trial[1]

Metabolite	Correlation (r)	P
γ-Tocopherol	−0.275	4.72×10^{-10}*
Methyl palmitate (15 or 2)	−0.187	2.82×10^{-5}*
Threonate	0.173	1.06×10^{-4}*
Pyridoxate	0.173	1.07×10^{-4}*
1-Arachidonoylglycero-phosphoethanolamine	−0.173	1.08×10^{-4}*
Pantothenate	0.170	1.42×10^{-4}
N-acetylalanine	−0.160	3.34×10^{-4}
17-Methylstearate	−0.155	5.14×10^{-4}
α-Tocopherol	0.153	6.46×10^{-4}
Hexanoylcarnitine	−0.152	6.80×10^{-4}
α-CEHC glucuronide	0.147	1.00×10^{-3}
1,7-Dimethylurate	−0.146	1.00×10^{-3}
Stearoyl sphingomyelin	−0.144	1.00×10^{-3}
1-Linoleoylglycero-phosphoethanolamine	−0.142	1.00×10^{-3}
1-Docosahexaenoylglycero-phosphocholine	0.142	2.00×10^{-3}
2-Arachidonoylglycero-phosphoethanolamine	−0.140	2.00×10^{-3}
cis-Vaccenate (18:1n–7)	−0.140	2.00×10^{-3}
3-Methoxytyrosine	−0.140	2.00×10^{-3}
Theobromine	−0.138	2.00×10^{-3}
Androsteroid monosulfate 1	−0.138	2.00×10^{-3}

[1] HEI was treated as a continuous variable. The strongest associations were selected by the smallest P value. Partial Pearson correlations adjusted for age, sex, smoking status (current smokers, former/never smokers), total energy intake (kcal/d), and multivitamin/supplement use (yes or no). The significance threshold was set at the Bonferroni-corrected level of $P < 1.214 \times 10^{-4}$ (0.05/412 known metabolites). Significant associations are indicated with an asterisk. HEI, Healthy Eating Index; PLCO, Prostate, Lung, Colorectal, and Ovarian; α-CEHC, 2,5,7,8-tetramethyl-2(2′-carboxyethyl)-6-hydroxychroman.

Implications for future study design

Although only 39 metabolites were conclusively determined to be associated with diet, there was clear evidence that a much larger percentage was correlated with usual dietary intake. The total number of metabolites associated with each food group at various FDR thresholds is shown in **Figure 1**. Approximately 130 metabolites associated with total alcohol consumption met the FDR threshold of 0.20, and ∼85 of these met the 0.05 threshold (Figure 1). Beer also had a large number of metabolite correlations, and liquor and coffee were correlated with >20 metabolites at the FDR threshold of 0.05.

The use of multiple samples from 30 individuals collected 1 y apart provided important information about the within-individual variability of metabolites (**Table 5**); the 1-y ICCs were variable but reasonable, with a range of 0.27–0.89. The median intra-assay CV, calculated by using replicate samples from a separate source of pooled serum, was 0.10 (IQR: 0.04–0.21).

To inform future study design, we used our data to determine the sample size that would be necessary to detect an association between metabolites and disease, which was measured as an RR. Considering dietary groups that had significant correlations with metabolites (Table 3), we then determined the sample size needed to have 80% power to detect an association in a 1:1 case-control study. For a large effect (RR = 3.0), sample sizes of ∼200–400 would be sufficient for most metabolites, whereas smaller effects (RR = 1.5) can only be detected in larger samples of ∼1100–3000 (Table 5). As expected, the number of individuals needed is smaller for metabolites with lower within-subject variability, and when possible, we should collect multiple biological specimens per subject to obtain a better estimate of usual metabolite levels. For example, to determine an association between the main citrus metabolite, stachydrine, and a health outcome and assuming an RR of 1.5, one would need a sample size of 2813 individuals; with a second specimen, the required sample size is reduced to 1898. Higher RRs can be detected with smaller samples; for stachydrine and an RR of 3.0, 398 individuals would be required with one specimen compared with 269 individuals with 2 specimens.

DISCUSSION

With the use of strict correction for multiple comparisons, the application of metabolomics to an epidemiologic data set detected 39 metabolites of known identity that were correlated with a total of 13 dietary groups. Metabolite levels were reproducible and stable over a year, indicating that metabolomics can be informative for nutritional epidemiologic studies. Moreover, the sample sizes needed to design an adequately powered study of metabolites and disease risk are realistically attainable.

Identification of biomarkers of diet

Our data replicated and validated findings from previous targeted biomarker studies, which supports the specificity and validity of our study. For example, previous studies also found correlations between citrus and stachydrine (29), which has been

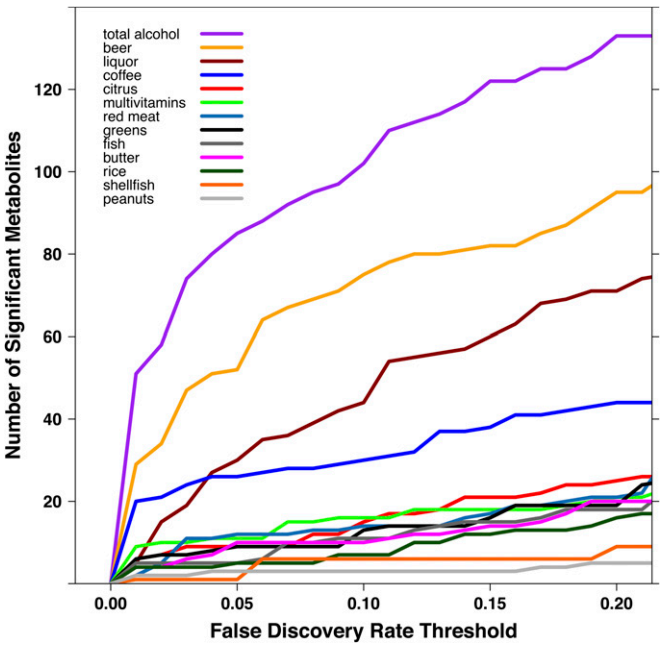

FIGURE 1. Many more metabolites are statistically associated with food groups by using the less conservative FDR method, as opposed to the Bonferroni method, for multiple testing correction. Only food groups with significant Bonferroni-corrected correlations are included. Each food group shown has at least one significant metabolite at an FDR of 0.01, and many more metabolites are found as the stringency of the FDR threshold is relaxed. FDR, false discovery rate.

TABLE 5
Sample size required to detect disease-metabolite associations for a case-control study[1]

				RR[4]		
Dietary group	Top metabolite[2]	One-year ICC[3]	No. of specimens	1.5	2.0	3.0
Citrus	Stachydrine	0.35	1	2813	999	398
			2	1898	674	269
Red meat	Indolepropionate[5]	0.81	1	1589	564	225
			2	1286	457	182
Fish	CMPF	0.33	1	2852	1012	404
			2	1917	681	272
Butter	Methyl palmitate (15 or 2)	0.62	1	1678	566	226
			2	1332	446	181
Peanuts	Tryptophan betaine	0.74	1	1345	478	191
			2	1163	413	165
Coffee	Trigonelline (*N′*-methylnicotinate)	0.74	1	1326	471	188
			2	1154	410	163
Beer	16-Hydroxypalmitate	0.42	1	1967	698	279
			2	1475	524	209
Liquor	Ethyl glucuronide	0.27	1	3528	1253	500
			2	2255	801	319
Multivitamin	Pantothenate	0.89	1	1245	442	176
			2	1114	395	158

[1] Total sample size, assuming 1:1 matching on case-control status. The metabolite with the strongest significant correlation was selected for each dietary group. Exceptions are for greens, shellfish, rice, and total alcohol because of the strongest metabolite correlation being shared in common (greens and shellfish both had CMPF as the most strongly correlated metabolite, which is already shown for fish; the most strongly correlated metabolite for total alcohol is ethyl glucuronide, which is already shown for liquor) and correlations that may reflect other foods (rice and DHA). Correlations were positive unless otherwise indicated; 80% power, $P = 0.05$. CMPF, 3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid (CMPF is also the top metabolite for green vegetables and shellfish); ICC, intraclass correlation coefficient.

[2] Top metabolite for partial Pearson correlation between dietary groups and metabolites; the lowest P value is the top metabolite.

[3] One-year ICC is a measure of the similarity between 2 specimens from the same individual with dates of blood collection separated by 1 y; $n = 30$ participants.

[4] RR compares risk for the fourth compared with the first quartile; true RR is measurement error corrected.

[5] Negative correlation between dietary group and metabolite.

identified as a biomarker of citrus in feeding studies (30) and urine metabolomics studies (10, 31). Scyllo- and chiro-inositol, also correlated with citrus intake in our study, have been previously associated with orange and citrus juice consumption (32). With regard to alcohol intake, several of the correlations we observed were lipid metabolites, which have previously been shown to contribute to an overall metabolomic profile of alcohol consumers (33). Some of the strongest associations we observed were for coffee components or downstream metabolites of these components (34, 35). Interestingly, we detected some correlations that may reflect co-consumption of other food items: for example, the association between DHA and rice, which may actually reflect co-consumption of rice with fish; DHA is an omega-3 fatty acid in fish oil and the correlations are of comparable strength.

In addition to identifying associations that have previously been reported in targeted biomarker studies, we identified many novel serum metabolite and diet correlations. The serum metabolites 4-vinylphenol sulfate and tryptophan betaine both reflected peanut consumption. The association of 4-vinylphenol sulfate, a xenobiotic involved in benzoate metabolism, with peanut consumption is plausible because it has previously been identified as a component in roasted peanuts (36). Tryptophan betaine, also known as hypaphorine, is an indole alkaloid that has been previously detected in legumes (37–39). A small study in breastfeeding mothers reported its presence in breast milk and also an association with peanut consumption (40). Moreover, many unnamed metabolites identified herein were also correlated with diet and therefore may be novel biomarkers; however, all food groups that were significantly correlated with unnamed metabolites were also correlated with identified metabolites.

Scoring higher on the HEI was negatively correlated with γ-tocopherol levels, which is consistent with the fact that γ-tocopherol is a major component of fried foods. Furthermore, a higher HEI score was positively associated with α-tocopherol, which is mainly found in healthier foods such as nuts, seeds, vegetable oils, and leafy greens. Even after adjustment for multivitamin/supplement use, most of the correlations with HEI were for vitamins or their constituents or metabolites, suggesting that these correlations are biologically plausible.

Our agnostic approach allowed the identification of novel associations and overcame the main challenges of traditional biomarker development, which typically necessitates the determination of biomarker kinetics, laboratory variability in measurement, and the capability of the biomarker signal to surpass laboratory error. By design, we identified biomarkers whose signals surpass this threshold. Further studies are needed to confirm the robust nature of our study design and the correlations we identified.

Implications for future studies

Our study used the Bonferroni method to correct for multiple comparisons, which reduces the probability of falsely identifying significant findings. Despite using this strict correction method,

multiple significant associations between serum metabolites and dietary variables were detected in our study. If we are willing to accept a higher error rate (FDR = 0.05), we would detect 1396 associations between metabolites and self-reported dietary variables (data not shown). Therefore, it is likely that there are many more true associations, but we lack the study power to detect them. The number of correlations by FDR cutoff, which are shown in Figure 1, provides a glimpse into the number of associations to be found with each food group.

We explored the variability in selected metabolites and the effect this may have on statistical power for planning future epidemiologic studies. The collection of multiple specimens per person would provide more precise measurements of the usual metabolite profile. In considering the use of metabolites in studies of disease, it is noted that large sample sizes are required, even among the metabolites most highly correlated with diet. Samples of this size, however, are attainable in epidemiologic cohorts. The reproducibility and relative stability of biomarker peak intensities, as assessed by the strong ICCs we observed, support the feasibility of applying metabolomics to epidemiologic data sets.

This study has many strengths, including the large sample size, richness of the dietary data, and large number of metabolites detected. Metabolomics analyses are an efficient use of biological samples, compared with candidate metabolite assays, when biospecimens are limited in large prospective studies. Biomarkers with correlations that cross our strict statistical significance threshold are, by their very nature, robust given that the signal is strong enough to be detected.

The main limitation of our study is that we compared a single serum sample with self-reported diet, and dietary questionnaires are known to result in substantial measurement error. However, measurement error in dietary intake would be expected to bias results toward the null, and we detected many significant correlations. Although the FFQ was extensive, we are limited in our ability to distinguish certain subtypes of foods that may be differentially associated with metabolites. Furthermore, we acknowledge that there are many ways to categorize dietary data. The categorization we used was primarily based on the USDA's MyPlate classification (14); we also considered subgroups based on proposed biological components of foods. The similarity in metabolites associated with dietary groups and their individual food components (data not shown) suggests that our conclusions are not dependent on the food group classification.

Our findings may not be generalizable because participants were largely white, and thus we may have missed correlations with foods specific to certain ethnic groups. The results we did observe, however, are highly generalizable, specific, and show strong associations. Any metabolomics study is inherently limited by the set of metabolites detected by a specific platform. We lacked information on absolute levels of serum metabolites because they were measured as peak intensities rather than as actual concentrations, which is a limitation. Last, variability in serum metabolites could be greatly influenced by the gut microbiota; for example, differences in metabolite levels may be attributable to differences in the gut microbiota rather than differences in dietary intake. The gut microbiota does change with age (41); however, it is thought to be relatively stable after early childhood (42). Although we were unable to assess the gut microbiome directly within this sample, there were no differences in the main findings by age.

In conclusion, the large number of correlations between self-reported diet and serum metabolites confirms that metabolomics can be applied to epidemiologic studies for identification of novel dietary biomarkers. There is a need for specific, reliable biomarkers that accurately reflect dietary intake and that can be applied to many populations. We emphasize, however, that although we appear to have uncovered objective biomarkers of diet, it should not yet be assumed that these biomarkers outperform self-report as a measure of usual dietary intake. Ultimately, whether a biomarker is a good measure of usual diet depends on the frequency of consumption of the food or nutrient, as well as the half-life of the metabolite. In addition, the identification of serologic metabolites not only reflects dietary intake but also metabolic processes, including the effects of genetic variation and the gut microbiota. Nevertheless, our metabolomic approach for identifying potential dietary biomarkers showed viable biomarkers for further investigation in feeding studies.

The authors' responsibilities were as follows— SCM, JNS, W-YH, RZS-S, RS, and AJC: designed the research; KAG, SCM, JNS, and AJC: conducted the research; SCM, JNS, W-YH, and AJC: provided essential materials; KAG, SCM, JNS, and QX: analyzed data; KAG, SCM, and AJC: wrote the manuscript; and KAG and AJC: had primary responsibility for final content. The authors had nothing to disclose and no conflicts of interest.

REFERENCES

1. Kaaks RJ. Biochemical markers as additional measurements in studies of the accuracy of dietary questionnaire measurements: conceptual issues. Am J Clin Nutr 1997;65(suppl):1232S–9S.
2. Ocké MC, Kaaks RJ. Biochemical markers as additional measurements in dietary validity studies: application of the method of triads with examples from the European Prospective Investigation into Cancer and Nutrition. Am J Clin Nutr 1997;65(suppl):1240S–5S.
3. Freedman LS, Schatzkin A, Wax Y. The impact of dietary measurement error on planning sample size required in a cohort study. Am J Epidemiol 1990;132:1185–95.
4. Jenab M, Slimani N, Bictash M, Ferrari P, Bingham SA. Biomarkers in nutritional epidemiology: applications, needs and new horizons. Hum Genet 2009;125:507–25.
5. Jones DP, Park Y, Ziegler TR. Nutritional metabolomics: progress in addressing complexity in diet and health. Annu Rev Nutr 2012;32:183–202.
6. Wild CP, Scalbert A, Herceg Z. Measuring the exposome: a powerful basis for evaluating environmental exposures and cancer risk. Environ Mol Mutagen 2013;54:480–99.
7. Beckmann M, Lloyd AJ, Haldar S, Favé G, Seal CJ, Brandt K, Mathers JC, Draper J. Dietary exposure biomarker-lead discovery based on metabolomics analysis of urine samples. Proc Nutr Soc 2013;72:352–61.
8. Lloyd AJ, Beckmann M, Haldar S, Seal C, Brandt K, Draper J. Data-driven strategy for the discovery of potential urinary biomarkers of habitual dietary exposure. Am J Clin Nutr 2013;97:377–89.
9. O'Sullivan A, Gibney MJ, Brennan L. Dietary intake patterns are reflected in metabolomic profiles: potential role in dietary assessment studies. Am J Clin Nutr 2011;93:314–21.
10. Pujos-Guillot E, Hubert J, Martin JF, Lyan B, Quintana M, Claude S, Chabanas B, Rothwell JA, Bennetau-Pelissero C, Scalbert A, 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 2013;12:1645–59.
11. Floegel A, von Ruesten A, Drogan D, Schulze MB, Prehn C, Adamski J, Pischon T, Boeing H. Variation of serum metabolites related to habitual diet: a targeted metabolomic approach in EPIC-Potsdam. Eur J Clin Nutr 2013;67:1100–8.
12. Prorok PC, Andriole GL, Bresalier RS, Buys SS, Chia D, Crawford ED, Fogel R, Gelmann EP, Gilbert F, Hasson MA, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Control Clin Trials 2000;21:273S–309S.

13. Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, McNutt S, McIntosh A, Rosenfeld S. Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. Am J Epidemiol 2001;154:1089–99.
14. USDA. Choose My Plate.Gov. Available from: www.choosemyplate.gov (cited 3 January 2014).
15. Tippett KS, Cypell YS. Design and operation: the Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey, 1994-96. Washington, DC: USDA, Agricultural Research Service, 1997:197. Nationwide Food Surveys Report No. 96-1.
16. Guenther PM, Casavale KO, Reedy J, Kirkpatrick SI, Hiza HA, Kuczynski KJ, Kahle LL, Krebs-Smith SM. Update of the Healthy Eating Index: HEI-2010. J Acad Nutr Diet 2013;113:569–80.
17. US Department of Health and Human Services; USDA. Dietary guidelines for Americans, 2010. 7th ed. Washington, DC: US Government Printing Office, 2010.
18. Evans AM, DeHaven CD, 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. Anal Chem 2009;81:6656–67.
19. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wägele B, Altmaier E, Deloukas P, Erdmann J, Grundberg E, et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature 2011;477:54–60.
20. Benjamini Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 1995;57:289–300.
21. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics. Metabolomics 2006;2:171–96.
22. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B 1996;58:267–88.
23. Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn 2003;52:239–81.
24. R Core Team. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2012.
25. Sampson JN, Boca SM, Shu XO, Stolzenberg-Solomon RZ, Matthews CE, Hsing AW, Tan YT, Ji BT, Chow WH, Cai Q, et al. Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. Cancer Epidemiol Biomarkers Prev 2013;22:631–40.
26. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, et al. HMDB 3.0: the Human Metabolome Database in 2013. Nucleic Acids Res 2013;41:D801–7.
27. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, et al. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 2009;37:D603–10.
28. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, et al. HMDB: the Human Metabolome Database. Nucleic Acids Res 2007;35:D521–6.
29. Chambers ST, Kunin CM. Isolation of glycine betaine and proline betaine from human urine: assessment of their role as osmoprotective agents for bacteria and the kidney. J Clin Invest 1987;79:731–7.
30. Atkinson W, Downer P, Lever M, Chambers ST, George PM. Effects of orange juice and proline betaine on glycine betaine and homocysteine in healthy male subjects. Eur J Nutr 2007;46:446–52.
31. Lloyd AJ, Beckmann M, Favé G, Mathers JC, Draper J. Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. Br J Nutr 2011;106:812–24.
32. Sanz ML, Villamiel M, Martínez-Castro I. Inositols and carbohydrates in different fresh fruit juices. Food Chem 2004;87:325–8.
33. Jaremek M, Yu Z, Mangino M, Mittelstrass K, Prehn C, Singmann P, Xu T, Dahmen N, Weinberger KM, Suhre K, et al. Alcohol-induced metabolomic differences in humans. Transl Psychiatry 2013;3:e276.
34. Niseteo T, Komes D, Belščak-Cvitanovićb A, Horžićb D, Budečc M. Bioactive composition and antioxidant potential of different commonly consumed coffee brews affected by their preparation technique and milk addition. Food Chem 2012;134:1870–7.
35. Consonni R, Cagliani LR, Cogliati C. NMR based geographical characterization of roasted coffee. Talanta 2012;88:420–6.
36. Walradt JP, Pittet AO, Kinlin TE, Muralidhara R, Sanderson A. Volatile components of roasted peanuts. J Agric Food Chem 1971;19:972–9.
37. Lu C-T, Tang H-F, Sun X-L, Wen A-D, Zhang W, Ma N. Indole alkaloids from chickpea seeds (Cicer arietinum L.). Biochem Syst Ecol 2010;38:441–3.
38. Hofinger M, Monseur X, Pais M, Jarreau FX. Further confirmation of the presence of indolylacrylic acid in lentil seedlings and identification of hypaphorine as its precursor. Phytochemistry 1975;14:475–7.
39. Tsopmo A, Muir AD. Chemical profiling of lentil (Lens culinaris Medik.) cultivars and isolation of compounds. J Agric Food Chem 2010;58:8715–21.
40. Keller BO, Wu BTF, Li SSJ, Monga V, Innis SM. Hypaphorine is present in human milk in association with consumption of legumes. J Agric Food Chem 2013;61:7654–60.
41. Arumugam M. Enterotypes of the human gut microbiome. Nature 2011;473:174–80.
42. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol 2011;9:279–90.