The Journal of Nutrition
Methodology and Mathematical Modeling

See corresponding editorial on page 2643.

Fecal Metabolites as Biomarkers for Predicting Food Intake by Healthy Adults

Leila M Shinn,[1] Aditya Mansharamani,[2] David J Baer,[3] Janet A Novotny,[3] Craig S Charron,[3] Naiman A Khan,[1,4] Ruoqing Zhu,[5,6] and Hannah D Holscher[1,4,6,7]

[1]Division of Nutritional Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [3]Beltsville Human Nutrition Research Center, Agricultural Research Service, USDA, Beltsville, MD, USA; [4]Department of Kinesiology & Community Health, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [5]Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [6]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA; and [7]Department of Food Science and Human Nutrition, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ABSTRACT

Background: The fecal metabolome is affected by diet and includes metabolites generated by human and microbial metabolism. Advances in -omics technologies and analytic approaches have allowed researchers to identify metabolites and better utilize large data sets to generate usable information. One promising aspect of these advancements is the ability to determine objective biomarkers of food intake.

Objectives: We aimed to utilize a multivariate, machine learning approach to identify metabolite biomarkers that accurately predict food intake.

Methods: Data were aggregated from 5 controlled feeding studies in adults that tested the impact of specific foods (almonds, avocados, broccoli, walnuts, barley, and oats) on the gastrointestinal microbiota. Fecal samples underwent GC-MS metabolomic analysis; 344 metabolites were detected in preintervention samples, whereas 307 metabolites were detected postintervention. After removing metabolites that were only detected in either pre- or postintervention and those undetectable in ≥80% of samples in all study groups, changes in 96 metabolites relative concentrations (treatment postintervention minus preintervention) were utilized in random forest models to *1*) examine the relation between food consumption and fecal metabolome changes and *2*) rank the fecal metabolites by their predictive power (i.e., feature importance score).

Results: Using the change in relative concentration of 96 fecal metabolites, 6 single-food random forest models for almond, avocado, broccoli, walnuts, whole-grain barley, and whole-grain oats revealed prediction accuracies between 47% and 89%. When comparing foods with one another, almond intake was differentiated from walnut intake with 91% classification accuracy.

Conclusions: Our findings reveal promise in utilizing fecal metabolites as objective complements to certain self-reported food intake estimates. Future research on other foods at different doses and dietary patterns is needed to identify biomarkers that can be applied in feeding study compliance and clinical settings. *J Nutr* 2022;152:2956–2965.

Keywords: gastrointestinal microbiota, metabolomics, fidelity measures, dietary intake biomarker, machine learning

Introduction

Traditionally, gut microbiome researchers have utilized DNA sequencing methods to characterize the composition of the gut microbiota, or "who" is there (1, 2). More recently, focus has shifted to the functionality of these microbes, leading to the use of metabolomics to discover and validate molecular by-products present in biological samples. This approach allows for the identification of cellular processes in response to stimuli, i.e., specific food consumption (3), because fecal samples contain human- and microbial-recovered metabolites, as well as by-products of nondigested and absorbed food components (4). One promising route for these discoveries is to complement self-reported measures of food intake and compliance with fecal metabolites as objective biomarkers. Although self-reported measures are frequently utilized in studies, their reliability and validity have been criticized owing to errors, including misreporting (5–9). Therefore, objective biomarkers that can complement self-reported measures of food intake are of enormous interest.

Manuscript received May 18, 2022. Initial review completed August 1, 2022. Revision accepted August 25, 2022.
First published online August 30, 2022; doi: https://doi.org/10.1093/jn/nxac195.

Researchers from multiple government agencies and public and private organizations have acknowledged the need to promote the discovery, development, and use of biomarkers across various applications (10–14). Metabolomic studies in the nutrition field have focused on specific metabolites associated with food consumption (15). For example, trimethylamine N-oxide (TMAO) has been identified as a potentially atherogenic gut-derived metabolite from dietary nutrients such as choline, betaine, and L-carnitine in eggs, red meat, and fish (16, 17). Other work has demonstrated that various nutrients can serve as food-specific biomarkers, including lutein (avocado), tocopherols (almond), proline betaine (citrus fruit), and methoxyeugenol glucuronide, dopamine sulfate, salsolinol sulfate, xanthurenic acid, and 6-hydroxy-1-methyl-1,2,3,4-tetrahydro-β-carboline sulfate (banana) (18–21). However, most identified biomarkers are from blood or urine samples, whereas the utility of fecal samples (a noninvasive biological sample) to generate biomarkers of food intake is underexplored (21–25). Our previous work demonstrated that fecal bacteria could be used to identify food intake with up to 85% accuracy (26). Although these exploratory efforts are not without their limitations, they reveal promise in pursuing noninvasive objective biomarker development through further exploration of the functional information available in fecal samples. One challenge under continued study is analyzing metabolomics data (3).

Thus, aligned with our previous effort (26), we aimed to develop a proof-of-concept machine learning model to identify fecal metabolites that could be leveraged as biomarkers of specific food intake. Herein, we describe secondary analyses conducted on data from fecal samples collected at pre- and postintervention of 5 feeding trials (almonds, avocados, broccoli, walnuts, and whole grains). The purpose of the present investigation was to utilize a computationally intensive, multivariate, machine learning approach to identify metabolite biomarkers that accurately predict food intake.

Methods

Experimental design

This study utilized data from 5 separate feeding studies examining almond (27), avocado (18, 28), broccoli (29), walnut (30), or whole-grain barley and whole-grain oat (31) consumption in adults ($n = 285$) between 21 and 75 y of age, which have been previously described. **Table 1** summarizes the study details briefly. **Supplemental Table 1** provides further details of the nutrient compositions of the provided meals. Briefly, the almond, broccoli, and walnut trials were complete feeding studies that utilized randomized, controlled, crossover designs. The whole-grain study was a 6-wk, complete feeding, randomized, controlled, parallel-arm design. The avocado trial was a randomized, controlled trial that provided 1 meal daily for 12 wk. All study procedures were administered in accordance with the Declaration of Helsinki and were approved by the Institutional Review Board of the MedStar Health Research Institute (almond, broccoli, walnut, and whole grains) or the University of Illinois Institutional Review Board (avocado).

Fecal metabolomics

Fecal samples were homogenized, divided into aliquots, and stored at −80°C until metabolomic analysis. Fecal extractions were performed at the Metabolomics Center, Roy J Carver Biotechnology Center, University of Illinois at Urbana-Champaign, using previously published protocols (32, 33). Briefly, two 1-mL fractions were taken from each sample and dried. One fraction was derivatized for 90 min at 500°C with 80 μL methoxyamine hydrochloride in pyridine (20 mg/mL), followed by a 60-min treatment at 500°C with 80 μL N-Methyl-N-trimethylsilyl-trifluoroacetamide (MSTFA). A 5-μL aliquot of a C31 fatty acid internal standard was added to each derivatized sample before trimethylsilylation (32, 33). Sample volumes of 1 mL were injected with a split ratio of 7:1 into a low-resolution GC-MS system (accuracy of 0.2 atomic mass units) consisting of an Agilent 7890A (Agilent Inc.) gas chromatograph, an Agilent 5975C mass selective detector, and an Agilent 7683B autosampler. The spectra of all chromatogram peaks were compared with electron impact mass spectrum libraries NIST08 [National Institute of Standards and Technology (NIST)], WILEY08 (Palisade Corporation), and a custom library of the University of Illinois metabolomics center. All data were normalized to the internal standard in each chromatogram to allow direct comparisons between samples. The chromatograms and mass spectra were evaluated using the MSD ChemStation (Agilent) and AMDIS (NIST).

Statistical analysis and validation

The raw preintervention data set included 344 metabolites, and the raw postintervention data set included 307 metabolites. See **Supplemental Figure 1** for additional details. We removed metabolites from the data set that were undetectable in ≥80% of samples in all study groups. We then assessed how the remaining 96 metabolites correlated with one another using Spearman rank correlation coefficients (34). A set of correlation coefficients were computed for each pair of metabolites across each food group's control arm only, and the weighted mean was taken to calculate the final correlation coefficient. Computing the correlations individually across the food groups was necessary to account for the differences in background diet.

For the following analyses, we then imputed the remaining missing values to be a random value between 0 and one-half of the minimum observed value across all observations (35). Furthermore, we considered the changes in relative concentration (treatment postintervention minus preintervention) for each fecal metabolite. Using relative concentration allows for the comparison of compounds across the batch of analyzed samples. It was measured as compound peak area/internal standard peak area and normalized to the sample weight. This normalization reduces deviation and makes the resulting data more uniform. Using the difference between preintervention and postintervention metabolites allowed for the determination of the internal differences, or batch effects (26), related to the performance of these 5 studies and their respective background diets. We then examined how the preintervention and postintervention differences in metabolites varied across each food group. We computed the mean difference for each metabolite within each food group's treatment and control and then plotted the log of the fold change ratio of the treatment group's mean difference for each metabolite with respect to the control group.

Next, we utilized random forest models to examine the relation between food consumption and change in fecal metabolome. For each group, a *scikit-learn* (36) random forest model with 5000 trees was trained using the difference in metabolites for that food group as the covariate and the participant's group (control or treatment) as the outcome. Because this effort focused on secondary analysis of data sets

Supported by the Foundation for Food and Agriculture Research New Innovator Award (to HDH), USDA National Institute of Food and Agriculture Hatch Project 1009249 (to HDH), the University of Illinois at Urbana-Champaign College of Agricultural, Consumer and Environmental Sciences Jonathan Baldwin Turner Fellowship (to LMS), USDA Agriculture and Food Research Initiative grant 1026383 (to LMS), and the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign through the NCSA Faculty Fellows program (to HDH and RZ).
Author disclosures: HDH and DJB have received research support from the USDA and the Almond Board of California. HDH and NAK have received research support from Hass Avocado Board. DJB has received research support from the California Walnut Commission and Kellog Company. LMS, AM, JAN, CSC, and RZ report no conflicts of interest. HDH is a member of *The Journal of Nutrition* Editorial Board.
Supplemental Tables 1–4 and Supplemental Figures 1–5 are available from the "Supplementary data" link in the online posting of the article and from the same link in the online table of contents at https://academic.oup.com/jn/.
LMS and AM contributed equally to this work.
Address correspondence to HDH (e-mail: hholsche@illinois.edu) or RZ (e-mail: rqzhu@illinois.edu).

TABLE 1 Study design of 5 studies in metabolically healthy adult participants aggregated for secondary analyses

	Population		Trial design		
Study	Age,[1] y	BMI,[1] kg/m^2	Design	Controlled diet composition	Intervention food
Almond ($n = 18$) (27)	57.0 ± 2.3 (25–75)	30.0 ± 1.0 (21.9–36.1)	Five 3-wk period randomized, controlled, crossover with 1-wk washouts	Complete feeding (55% carbohydrate, 15% protein, 30% fat)	1.5 servings (42 g) per day of roasted, chopped almonds (base diet scaled down for isocaloric inclusion)
Avocado ($n = 163$) (28)	35.0 ± 0.5 (25–45)	32.8 ± 0.5 (23.9–58.8)	12-wk randomized, controlled, parallel-arm	One daily meal (45% carbohydrate, 15% protein, 40% fat)	175 g (males) or 140 g (females) avocado (once-daily isocaloric meal)
Broccoli ($n = 18$) (29)	55.0 ± 1.7 (21–70)	28.0 ± 1.2 (19.0–36.6)	Two 18-d period randomized, controlled, crossover with 24-d washout period	Complete feeding (54% carbohydrate, 16% protein, 30% fat)	200 g cooked broccoli with 20 g raw daikon radish per day (added to controlled diet)
Walnut ($n = 18$) (30)	53.1 ± 2.2 (25–75)	28.8 ± 0.9 (20.2–34.9)	Two 3-wk period randomized, controlled, crossover with 1-wk washouts	Complete feeding (54% carbohydrate, 17% protein, 29% fat)	1.5 servings (42 g) per day of walnuts (base diet scaled down for isocaloric inclusion)
Whole grains ($n = 68$) (31)	52.8 ± 1.3 (25–70)	28.2 ± 0.5 (18.9–38.3)	6-wk randomized, controlled, parallel-arm	Complete feeding with 0.7 servings (11.2 g) of whole grains per 1800 kcal (53% carbohydrate, 15% protein, 32% fat)	4 servings (64 g) of whole-grain *1*) barley or *2*) oats per 1800 kcal

[1] Values are mean ± SE (range).

from 5 different studies, we needed to consider that the foods were from 5 sample populations. To address this, the contrast between the pre- and postintervention of each study population was used to distinguish different food effects. Leave-one-out cross-validation was the most appropriate computational method to address this change, and the lack of a separate validation data set (37). Thus, the optimal random forest classifier parameters were found via a leave-one-out grid search using each model's out-of-bag score as the search metric. The variable training parameters in the search included *max_features*, the number of features to consider when looking for the best split, and *min_samples_leaf*, the minimum number of samples required to be at a leaf node. All other training parameters were fixed at their default values. After the optimal parameters were found, each model was trained and evaluated in a leave-one-out cross-validated fashion.

Finally, we used random forests to model the relation between food consumption and change in fecal metabolome across multiple food groups. Food groups were only included in the multifood models if the random forest model built for that single food in the previous step could accurately separate the treatment and control arms for that specific study. The baseline score (i.e., randomly guessing) for a binary classification problem is 50%; thus, only food groups that achieved significantly higher performance than this baseline score would be included. The treatment metabolite data set (consisting only of samples in which the study participant consumed the food intervention) was used as the covariate, and the food consumed was used as the outcome for a *scikit-learn* random forest classifier with 5000 trees. The model was trained and evaluated in the same fashion as previously described.

We quantified the batch effect to verify that the random forest model was learning relations based on the food consumed rather than participation in a specific study. We consider the shift in the fecal metabolites before (pre-) and after (post-) the intervention in the treatment groups (i.e., the treatment signal) to be driven by 2 distinct effects, as previously described (26). The first is the effect of the food (e.g., almond) or treatment effect. The second is the effect of the background diet (the batch effect), which indicates all the foods each participant consumed aside from the study-specific food of interest. We considered any metabolomic signatures present in both control and treatment participants as a batch effect because their presence could be learned by the random forest, thereby artificially inflating the model classification performance. After training and evaluating the multifood model on the treatment metabolite data set, the same model was evaluated on the control metabolite data set to quantify the batch effect. High classification performance on the control metabolite data set would indicate that the model was exploiting metabolomic signatures present in both control and treatment participants, i.e., the model was predictive of study participation, not food intake. For the models evaluating the food compared with their control only (i.e., the single-food models), the random forest models trained on each study's treatment and control data are not susceptible to the batch effect, because the models are not comparing treatment groups across different study settings. Thus, we did not quantify the batch effect for the single-food models.

In addition to investigating whether food intake could be predicted by changes in fecal metabolome composition, a second primary aim of the study was to identify a compact set of metabolites that could be used as food intake biomarkers. Thus, we examined the feature importance scores produced by the random forest models to rank the fecal metabolites according to their predictive power. Highly ranked features from the single-food random forest models would indicate metabolites that show promise in acting as biomarkers for specific food intake. Highly ranked features from the multifood random forest could further reveal potential biomarkers unique to specific foods.

Results

The raw metabolomic data set contained 372 observations for the preintervention and postintervention samples for all study participants across the 5 feeding studies (Supplemental Figure 1). The preintervention data set included 344 metabolites, and the postintervention data set included 307 metabolites. After participants that were missing either pre- or post- data were removed, 362 observations remained. Metabolites that were only detected in either the pre- or post- data were excluded. Metabolites that were undetectable in ≥80% of samples in all study groups were removed, with the remaining values imputed between 0 and one-half of the minimum observed value across all observations. A total of 96 metabolites remained after data preprocessing. **Supplemental Figure 2** shows a heat map of Spearman correlations between relative concentrations of these 96 metabolites across treatment and control groups in almond, avocado, broccoli, walnuts, and whole grains. Finally, the differences between pre- and post- relative

FIGURE 1 Heat maps of log-fold change between mean differences of relative concentrations of fecal metabolites in metabolically healthy adult participants consuming almond, avocado, broccoli, walnuts, and whole grains. (A) Lipids and lipid-like molecules, (B) organoheterocyclic compounds, (C) other organic acids and derivatives, (D) organic oxygen compounds, (E) amino acids, and (F) other metabolites. Postintervention metabolites were subtracted from the preintervention data to compute the net effect of the control and treatment interventions for each participant ($n = 181$). Log-fold change ratios were then computed for each metabolite's mean difference in each treatment group with respect to the corresponding control group. Orange boxes indicate an increased fold change from pre- to postintervention, whereas blue boxes indicate a decreased fold change. The darker the color, the higher the magnitude of change is for that metabolite. The dendrogram (black bars) was generated using a Euclidean distance metric for both study groups and the individual metabolites. Bars across the top and *y* axis show how variables cluster together. Items that are in the same cluster are more similar (i.e., across the top, hierarchical clusters show which foods have similar patterns of fold change across the metabolite; across the *y* axis, the clusters show which metabolites have similar patterns of fold change across the food groups).

metabolite concentrations for those 362 observations were calculated to generate 181 observations for subsequent analysis using leave-one-out cross-validation. These 181 observations made up the treatment data set (n = 103), i.e., the study periods where study diets included the specific foods, and the control data set (n = 78). **Figure 1** is a heat map of fold changes in metabolites between treatment and control groups.

Single-food models

Using the change in relative concentration of the 96 metabolites, the 6 single-food random forest models for almond, avocado, broccoli, walnuts, whole-grain barley, and whole-grain oats had prediction accuracies of 82%, 59%, 47%, 89%, 57%, and 55%, respectively (**Table 2**). The avocado, broccoli, whole-grain barley, and whole-grain oats prediction models performed poorly, with prediction scores near a baseline of 50%. Thus, these foods were not included in our multifood random forest. **Supplemental Table 2** reports the metabolites with the top 10 feature importance scores extracted from each single-food random forest model. The distribution of feature importance scores varied across the random forest models. The almond, avocado, whole-grain barley, and whole-grain oats models' feature importance scores were largely dominated by their top feature. For example, in the almond model, 10-hydroxystearic acid had an importance of 0.51 compared with linoleic acid (importance = 0.084), the next feature selected by the model. **Supplemental Figure 3** compares the control and almond groups' changes in relative concentrations of 10-hydroxylstearic acid (Supplemental Figure 3A) and linoleic acid (Supplemental Figure 3B). On the other hand, the broccoli and walnut models did not demonstrate similar steep drop-offs in variable importance. For example, in the walnut data, uric acid and 5-hydroxyindole-3-acetic acid had feature importance scores of 0.101 and 0.097, respectively. **Supplemental Figure 4** compares the control and walnut groups' change in relative concentrations of 5-hydroxyindole-3-acetic acid (Supplemental Figure 4A) and uric acid (Supplemental Figure 4B). It is important to note that the variable importance scores assigned by the random forest model are numerically unstable and may change slightly each time the model is refit. This shift occurs because of nondeterminism intrinsic to the random forest algorithm (38).

Mixed-food model

Owing to the poor performance of the other single-food random forest models, only the almond and walnut groups were included in the mixed-food random forest model, differentiating almond from walnut intake. The overall mixed-food random forest model classification accuracy was 91% (**Table 3**). As an additional validation step for the model, we examined respective control data independently to ensure we were truly measuring the differential impacts of the foods consumed rather than participation in a specific study. Thus, we used the mixed-food random forest to classify the control data set to ensure the model's training had not been influenced by batch effects (26). When the control data set was evaluated, the almond and walnut control mixed-food model appropriately had poor predictive accuracy (47% accuracy), indicating that the model was capable of distinguishing between almond and walnut consumption and not merely learning to separate participation in the almond or the walnut study. Their top feature largely dominated the mixed-food (almond and walnut) model's feature importance scores. The most important feature, 5-hydroxyindole-3-acetic acid (importance = 0.77), was almost 20 times more important than the second most important feature, α-tocopherol (importance = 0.042). **Supplemental Figure 5** compares the almond and walnut groups' change in relative concentrations of 5-hydroxyindole-3-acetic acid (Supplemental Figure 5A) and α-tocopherol (Supplemental Figure 5B). **Supplemental Table 3** reports the top 10 feature importance scores extracted from the almond and walnut mixed-food model.

Random forest classification

In practice, random forests tend not to assign groups of correlated features high feature importance scores. Although α-tocopherol received a low variable importance score from the single-food almond model, its mean difference was well-correlated with 10-hydroxystearic acid (r^2 = 0.42) and weakly correlated with linoleic acid (r^2 = −0.20), the top 2 important metabolites for the single-food almond model (Supplemental Figure 3). The difference in the relative concentrations of these 2 single-food almond model metabolites may be just as effective a differentiator, or a more effective differentiator, for almond consumption when compared with the difference in the relative concentration of α-tocopherol, therefore warranting a lower feature importance score for the latter. **Supplemental Table 4** provides Pearson correlation coefficients between the highly ranked metabolites in the almond single-food, walnut single-food, and mixed-food models.

Pooling treatment group data from all 6 food groups for classification via random forest demonstrated high performance [overall accuracy = 78%, receiver operating characteristic ROC, AUC = 0.95]; however, these results were likely due to batch effects. Specifically, we observed that whereas the single-food random forest models for avocado, broccoli, whole-grain barley, and whole-grain oats could not differentiate between participants in their respective treatment and control arms, the mixed-food random forest could differentiate these foods. Furthermore, when the mixed-food random forest model was used to classify the control participants across each study, it achieved 51% accuracy overall and high accuracy within the avocado (76%) and broccoli (87%) control groups. These results indicate that the mixed-food random forest's performance was derived largely from identifying relations between the metabolite changes and the study (i.e., the batch effect) rather than the food each study participant consumed. In other words, without the batch effect, the mixed-food model would have had lower performance when classifying control participants. Methods to remove the batch effect have been described previously (26); however, these methods may also remove additional non–batch effect signals from the data set, which would only further reduce classification performance. Because we observed low classification accuracy in the single-food models for avocado, broccoli, whole-grain barley, and whole-grain oats, removing the batch effect from these data and rebuilding the random forest models would only serve to decrease the accuracy of these single-food models and, as such, they would still not warrant inclusion in the multifood model. Thus, removing the batch effect would only be a viable strategy if we observed high performance of the single-food models and desired to verify that this high performance was truly derived from whole food consumption. For this analysis,

TABLE 2 Prediction of specific food intake in metabolically healthy adult participants using random forest compared with respective control groups

	Predicted label		
True label	Treatment, *n*	Control, *n*	Accuracy, %
	Almond, *n*	Control, *n*	
Almond, *n*	10	4	71
Control, *n*	1	13	93
Overall accuracy, %			82
	Avocado, *n*	Control, *n*	
Avocado, *n*	21	6	78
Control, *n*	12	5	29
Overall accuracy, %			59
	Barley, *n*	Control, *n*	
Barley, *n*	8	6	57
Control, *n*	6	8	57
Overall accuracy, %			57
	Broccoli, *n*	Control, *n*	
Broccoli, *n*	6	9	40
Control, *n*	7	8	53
Overall accuracy, %			47
	Oats, *n*	Control, *n*	
Oats, *n*	8	7	53
Control, *n*	6	8	57
Overall accuracy, %			55
	Walnut, *n*	Control, *n*	
Walnut, *n*	17	1	94
Control, *n*	3	15	83
Overall accuracy, %			89

larger sample sizes and fixed background diets across study groups would be needed to best address the presence of batch effects.

Discussion

Herein, we report fecal metabolites associated with individual food intake (i.e., almond, avocado, broccoli, whole-grain oat, whole-grain barley, and walnut). This effort, which utilized random forest to identify food intake biomarkers, revealed high predictive accuracy of almond and walnut intake, both in a single- (compared with respective controls) and in a mixed-food model (almond compared with walnut). These findings establish the potential role of fecal metabolites to objectively complement self-reported food measures and study compliance.

Random forest models can effectively classify and select biomarkers in metabolomics data (38). For example, Asnicar et al. (39) examined links between habitual diet and the microbiome using random forest models, after training on quantitative microbiome features, to predict dietary variables from FFQs. Although larger data sets are generally used in machine learning models, random forest models have effectively classified data sets with smaller sample sizes (40,41). Unlike some other supervised machine learning models (such as logistic regression and support vector machines), random forests easily generalize from binary to multiclass problems. Also, random forests have fewer risks of overfitting than support vector machines models and can handle highly collinear metabolomics data, unlike linear methods such as orthogonal projection to latent structure/partial linear regression (42). Further, random forests can intrinsically inform biomarker discovery by assigning importance scores to input features without relying on external feature selection tools. Because a high score indicates the metabolite was useful in classifying the food, metabolites with high feature importance scores could be promising biomarker candidates. Finally, because of the exploratory nature of this work, we utilized leave-one-out cross-validation on our 2 time point data set to statistically mimic the validation error (37).

Our mixed-model results revealed that only the almond and walnut groups demonstrated high classification performance when compared with their respective control groups and against one another. Of interest, participants in both the almond and

TABLE 3 Prediction of almond compared with walnut intake in metabolically healthy adult participants using random forest

	Predicted label		
True label	Almond, *n*	Walnut, *n*	Accuracy, %
Almond, *n*	13	1	93
Walnut, *n*	2	16	89
Overall accuracy, %			91

walnut controlled-feeding, randomized, crossover trials were provided with identical background diets, with the exception of almond or walnut supplementation (27, 30). This alludes to the importance of well-designed randomized clinical trials to study the impact of diet on the fecal microbiome.

For almond, 10-hydroxystearic acid (18:0) and linoleic acid (18:2n–6, cis-9,12) were identified as important features for differentiating treatment from control by our model. Of note, extracts from 8 Californian almond cultivars had a fatty acid profile ranging from 57% to 74% oleic acid (18:1n–9), and from 19% to 35% linoleic acid, with small amounts of α-linolenic acid (18:3n–3) (43). The appearance of 10-hydroxystearic acid in our samples is likely due to bacterial metabolism because some bacterial proteins possess oleate hydratases that catalyze the hydration or isomerization of double bonds in unsaturated fatty acids (44, 45).

For walnut, hydroxyindole-3-acetic acid (5-HIAA) and uric acid were important features for differentiating treatment from control samples. A previous review completed as part of the Food Biomarkers Alliance (FoodBAll) consortium identified 5-HIAA, a tryptophan and serotonin pathway metabolite, as a promising biomarker for walnut intake (25) because previous studies have associated walnut consumption with increased 5-HIAA in the urine (46–50) and serum (51). Therefore, our novel finding extends this work by reporting that 5-HIAA is a useful fecal biomarker for walnut intake.

Two metabolites, 5-HIAA (almond decreased compared with walnut) and α-tocopherol (walnut decreased compared with almond), differentiated almond from walnut consumption in our study (Supplemental Figure 5). (All changes in metabolite levels are reported as mean relative concentrations per 100 mg fecal weight.) 5-HIAA is a microbially derived fecal fermentation end product (52–54), and α-tocopherol has been established as a compliance measure for almond intake using blood samples (19, 56–58). Almonds and walnuts provide 89 mg and 71 mg tryptophan per 42-g serving, respectively; almonds provide 19 mg α-tocopherol/100 g almonds, whereas walnuts provide 0.9 mg α-tocopherol/100 g walnuts (59). Although α-tocopherol did not appear as a top feature in our single-food almond model using the optimal random forest parameters, the difference between α-tocopherol in the almond treatment and in the control was 1000% higher. Conversely, the mean difference in relative concentration from pre- to postintervention of α-tocopherol for the walnut treatment in our study was 100% lower. This may be due in part to maldigestion of nutrients contained in the plant cell walls (27, 30). Tocopherols have been shown to be only 11%–51% bioaccessible in peanuts and tree nuts when assessed using an in vitro digestion method (60).

In our previous work (26), collapsing whole-grain oats and whole-grain barley into a single "whole grains" category improved classification performance. However, herein, the combined whole grains random forest model did not improve accuracy over the single-food whole-grain barley (57%) and oats (55%) models, possibly owing to a weak signal or the parallel-arm design. This result highlights the ability of complete-feeding, randomized, crossover trials to control for confounding. This is further supported by the poor classification of the avocado group (59%), the other parallel-arm design trial in this study sample (18, 28). Interestingly, although the broccoli study was a controlled, complete-feeding, randomized, crossover study, our classification performance was low (47%). Our previous work utilizing microbiota (16S) data to identify fecal microbes as food intake biomarkers also had the worst performance with broccoli (26). The small sample size ($n = 15$ participants; $n = 30$ samples) available for metabolomics analysis or the similar nutrient composition to other foods under investigation may have contributed to the metabolic signature being inadequate to classify broccoli intake in the multifood model.

Although the current effort is not without limitations, it does provide important insights into the potential of modeling metabolomic data sets to determine food intake biomarkers. Further, it highlights the need for future research of rigorous design with metabolomic endpoints as a primary outcome. Of note, the 5 nutrition studies utilized herein were powered for their respective primary outcomes (none of which were fecal metabolomic analyses). One could expect that adequately powered future work that includes metabolomic endpoints as primary outcomes will further validate fecal metabolite biomarkers of food intake but will likely necessitate larger sample sizes, a factor that will come with increased cost. Our results also highlight important study design considerations. For example, the avocado and whole grain studies were conducted using parallel-arm designs, which may have contributed to the lack of signal in those single-food models. The intervention approach (i.e., single meal compared with complete feeding) is another important consideration—the low accuracy of the avocado single-food model was likely also affected by the feeding intervention being less controlled. The importance of controlling the background diet to detect fecal metabolite biomarkers is best illustrated by the predictive accuracy of the multifood model (almond compared with walnut)—although almonds and walnuts have the most similar nutrient compositions of the 5 foods, the use of identical background diets in both studies allowed these foods to be correctly differentiated from one another using fecal metabolites with 91% accuracy. Thus, one can hypothesize that if all 5 studies had been complete feeding, crossover trials with identical control and background diets (outside of the treatment foods), we would expect to have seen a stronger signal for the other foods. In an ideal trial, the same participants would have served as their own control and gone through each of the intervention arms (almond, avocado, broccoli, walnut, whole-grain barley, and whole-grain oats) with appropriate washout periods between each condition; however, trade-offs for that robust design would include increased participant fatigue and costs due to the extended study duration. That is not to say that with appropriate primary endpoints and large enough sample sizes, parallel-arm studies could not be used to identify fecal metabolite biomarkers. In fact, parallel-arm studies can prevent carryover effects between conditions that can affect causal inferences (61, 62). In summary, the current effort reveals promise in using fecal metabolites as biomarkers of food intake. Still, future work should examine these outcomes as primary endpoints using an appropriate study design.

Fecal microbial signatures associated with specific foods and nutrients have recently appeared in the literature (26, 39, 63, 64). Similar to recent work, the metabolites identified by our analyses are dominated by lipids and amino acids (65, 66). Although most efforts have utilized the blood or urine metabolome, a recent meta-analysis identified 273 fecal metabolites in 9 healthy data sets containing 779 samples from 629 individuals (65). Other large-scale studies also demonstrate the feasibility of collecting fecal samples in human studies. For example, the American Gut Project (66) and Twins UK (67) have amassed >15,000 and >5000 samples for microbiome analysis, respectively. However, there are limitations related to collecting fecal samples that should be acknowledged. For one, the cost

of metabolomic analyses is a barrier. However, as researchers continue to unveil the utility of metabolomics, the costs may drastically decrease, as seen with genome sequencing over the past quarter-century (68). Also, requesting study participants collect their fecal samples for research may contribute to selection bias. Therefore, the underrepresentation of specific communities should be addressed with targeted and population-based studies in the future (67).

Although the current effort focuses on the fecal metabolome, there is also a need for continued elucidation of the food metabolome to better delineate how dietary intake affects the fecal metabolome. With >25,000 known food compounds, the food metabolome presents an important source of novel dietary biomarkers. Most current dietary biomarkers are based on known food compositions and hypothesis-driven approaches, but metabolomics has facilitated the identification of novel biomarkers in various foods (69). These discoveries have led to databases such as FooDB (70), Phenol-Explorer (71), PhytoHub (72), and KNApSAcK (73), which harbor important information on endogenous, microbial, biotransformed, and exogenous/xenobiotic compounds in foods. Understanding the impact of the food metabolome on microbial metabolism and the endogenous human metabolome (or host metabolites), and subsequent fecal metabolites is essential for understanding the biotransformations from ingestion to excretion. Continued validation of food biomarkers, new metabolite identifications, and multiomics integration will move current black-box approaches toward understanding underlying mechanisms to advance nutritional epidemiology (74).

Future works should also continue to advance the field of molecular nutrition. In other words, there is an understanding of nutrient metabolism, and efforts such as the current study provide insight into the end products of this metabolism, i.e., fecal metabolome. However, as researchers continue to study nutritional end products, understanding the metabolome of the food products entering the digestive tract must also be captured. Consequently, comparisons from the start to the end of digestion can be made to further understand the bacterial biotransformations occurring throughout the digestive process. Reflecting on our lack of signal from our avocado, broccoli, and whole grains trials, the importance of future research using robust study designs with metabolomic analyses as a primary aim is clear. Future randomized, controlled, complete feeding trials should also examine the dose-response of specific foods and expand into other foods and dietary patterns. Dose-response studies are crucial because we must consider varying amounts of foods consumed to utilize the identified biomarkers in future observational studies. Furthermore, the reproducibility and cross-comparison of these works can be enhanced by developing reference materials (14). Once strongly designed research and continued development of databases support biomarkers of food intake, these measures can be studied in observational trials and later utilized in clinical and research settings as compliance measures to complement self-reported measures of intake.

In summary, using metabolomics data and machine learning, we have revealed promise in the feasibility of fecal metabolites as objective biomarkers of food intake by healthy adults. These findings provide groundwork for uncovering additional biomarkers of food intake. With future work and expansion to other foods and dietary patterns, biomarkers like the ones identified in this effort can be applied in feeding study compliance and clinical settings.

Acknowledgments

We thank Zhong (Lucas) Li (now at Duke Proteomics and Metabolomics Shared Resource, Center for Genomic and Computational Biology, Duke University School of Medicine) and Alexander Ulanov of the Metabolomics Center, Roy J Carver Biotechnology Center, University of Illinois at Urbana-Champaign for fecal extractions, GC-MS, and library comparisons on the completion of our metabolomics analyses. Further, we thank Melisa Bailey, Heather Guetterman, Jennifer (Kaczmarek) Burton, Annemarie Mysonhimer, Andrew Taylor, and Sharon Thompson for their technical assistance with processing the fecal samples in the primary studies. The authors' responsibilities were as follows—DJB, JAN, CSC, NAK, HDH, and RZ: designed the research; LMS: conducted the research; AM: analyzed the data and performed the statistical analysis; LMS and AM: wrote the first draft of the paper; HDH and RZ: critically reviewed and edited the manuscript and had primary responsibility for the final content; and all authors: read and approved the final manuscript.

Data Availability

The anonymized data and all of the code used for the analyses can be accessed from Github at https://github.com/holscher-nhml/usda-path-metabolomics.

References

1. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012;486(7402):207–14.
2. Claesson MJ, Clooney AG, O'Toole PW. A clinician's guide to microbiome analysis. Nat Rev Gastroenterol Hepatol 2017;14(10):585–95.
3. Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV, Jarmusch AK, Dorrestein PC. Mass spectrometry-based metabolomics in microbiome investigations. Nat Rev Microbiol 2022;20(3):143–60.
4. Zhgun ES, Ilina EN. Fecal metabolites as non-invasive biomarkers of gut diseases. Acta Naturae 2020;12(2):4–14.
5. Schatzkin A, Subar AF, Moore S, Park Y, Potischman N, Thompson FE, et al. Observational epidemiologic studies of nutrition and cancer: the next generation (with better observation). Cancer Epidemiol Biomarkers Prev 2009;18(4):1026–32.
6. Freedman L, Potischman N, Kipnis V, Midthune D, Schatzkin A, Thompson F, et al. A comparison of two dietary instruments for evaluating the fat–breast cancer relationship. Int J Epidemiol 2006;35(4):1011–21.
7. Rennie K, Coward A, Jebb S. Estimating under-reporting of energy intake in dietary surveys using an individualised method. Br J Nutr 2007;97(6):1169–76.
8. Poslusna K, Ruprich J, de Vries JHM, Jakubikova M, van't Veer P. Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice. Br J Nutr 2009;101(Suppl 2):S73–85.
9. Kipnis V, Midthune D, Freedman L, Bingham S, Day N, Riboli E, et al. Bias in dietary-report instruments and its implications for nutritional epidemiology. Public Health Nutr 2002;5(6a):915–23.
10. Meyers LD, Suitor CW. Dietary reference intakes research synthesis: workshop summary. Washington (DC): National Academies Press; 2007.
11. Raiten DJ, Namasté S, Brabin B, Combs GJ, L'Abbe MR, Wasantwisut E, et al. Executive summary—Biomarkers of Nutrition for Development: building a consensus. Am J Clin Nutr 2011;94(2):633S–50S.
12. Maruvada P, Lampe JW, Wishart DS, Barupal D, Chester DN, Dodd D, et al. Perspective: dietary biomarkers of intake and exposure—exploration with omics approaches. Adv Nutr 2020;11(2):200–15.

13. Nogal B, Blumberg JB, Blander G, Jorge M. Gut microbiota–informed precision nutrition in the generally healthy individual: are we there yet? Curr Dev Nutr 2021;5(9):nzab107.
14. Mandal R, Cano R, Davis CD, Hayashi D, Jackson SA, Jones CM, et al. Workshop report: toward the development of a human whole stool reference material for metabolomic and metagenomic gut microbiome measurements. Metabolomics 2020;16(11):119.
15. Rafiq T, Azab SM, Teo KK, Thabane L, Anand SS, Morrison KM, et al. Nutritional metabolomics and the classification of dietary biomarker candidates: a critical review. Adv Nutr 2021;12(6):2333–57.
16. Papandreou C, Moré M, Bellamine A. Trimethylamine N-oxide in relation to cardiometabolic health—cause or effect? Nutrients 2020;12(5):1330.
17. Wang Z, Klipfell E, Bennett B, Koeth R, Levison B, Dugar B, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 2011;472(7341):57–63.
18. Edwards CG, Walk AM, Thompson SV, Reeser GE, Erdman JW, Burd NA, et al. Effects of 12-week avocado consumption on cognitive function among adults with overweight and obesity. Int J Psychophysiol 2020;148:13–24.
19. Tan SY, Mattes RD. Appetitive, dietary and health effects of almonds consumed with meals or as snacks: a randomized, controlled trial. Eur J Clin Nutr 2013;67(11):1205–14.
20. Heinzmann SS, Brown IJ, Chan Q, Bictash M, Dumas M-E, Kochhar S, et al. Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. Am J Clin Nutr 2010;92(2):436–43.
21. Vázquez-Manjarrez N, Weinert CH, Ulaszewska MM, Mack CI, Micheau P, Pétéra M, et al. Discovery and validation of banana intake biomarkers using untargeted metabolomics in human intervention and cross-sectional studies. J Nutr 2019;149(10):1701–13.
22. Sri Harsha PSC, Wahab RA, Cuparencu C, Dragsted LO, Brennan L. A metabolomics approach to the identification of urinary biomarkers of pea intake. Nutrients 2018;10(12):1911.
23. Woodside JV, Draper J, Lloyd A, McKinley MC. Use of biomarkers to assess fruit and vegetable intake. Proc Nutr Soc 2017;76(3):308–15.
24. Münger LH, Garcia-Aloy M, Vázquez-Fresno R, Gille D, Rosana ARR, Passerini A, et al. Biomarker of food intake for assessing the consumption of dairy and egg products. Genes Nutr 2018;13(1):26.
25. Garcia-Aloy M, Hulshof PJM, Estruel-Amades S, Osté MCJ, Lankinen M, Geleijnse JM, et al. Biomarkers of food intake for nuts and vegetable oils: an extensive literature search. Genes Nutr 2019;14(1):7.
26. Shinn LM, Li Y, Mansharamani A, Auvil LS, Welge ME, Bushell C, et al. Fecal bacteria as biomarkers for predicting food intake in healthy adults. J Nutr 2021;151(2):423–33.
27. Novotny JA, Gebauer SK, Baer DJ. Discrepancy between the Atwater factor predicted and empirically measured energy values of almonds in human diets. Am J Clin Nutr 2012;96(2):296–301.
28. Thompson SV, Bailey MA, Taylor AM, Kaczmarek JL, Mysonhimer AR, Edwards CG, et al. Avocado consumption alters gastrointestinal bacteria abundance and microbial metabolite concentrations among adults with overweight or obesity: a randomized controlled trial. J Nutr 2021;151(4):753–62.
29. Charron CS, Vinyard BT, Ross SA, Seifried HE, Jeffery EH, Novotny JA. Absorption and metabolism of isothiocyanates formed from broccoli glucosinolates: effects of BMI and daily consumption in a randomised clinical trial. Br J Nutr 2018;120(12):1370–9.
30. Baer DJ, Gebauer SK, Novotny JA. Walnuts consumed by healthy adults provide less available energy than predicted by the Atwater factors. J Nutr 2016;146(1):9–13.
31. Thompson SV, Swanson KS, Novotny JA, Baer DJ, Holscher HD. Gastrointestinal microbial changes following whole grain barley and oat consumption in healthy men and women. FASEB J 2016;30(S1):406.1.
32. Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L. Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. Plant J 2000;23(1):131–42.
33. Braundmeier-Fleming A, Russell NT, Yang W, Nas MY, Yaggie RE, Berry M, et al. Stool-based biomarkers of interstitial cystitis/bladder pain syndrome. Sci Rep 2016;6(1):26083.
34. Spearman C. Demonstration of formulae for true measurement of correlation. Am J Psychol 1907;18(2):161–9.
35. Borgogna J-LC, Shardell MD, Yeoman CJ, Ghanem KG, Kadriu H, Ulanov AV, et al. The association of *Chlamydia trachomatis* and *Mycoplasma genitalium* infection with the vaginal metabolome. Sci Rep 2020;10(1):3420.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
37. Wang H, Yang F, Luo Z. An experimental study of the intrinsic stability of random forest variable importance measures. BMC Bioinform 2016;17:60.
38. Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, et al. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evid Based Complement Alternat Med 2013: 298183.
39. Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat Med 2021;27(2): 321–32.
40. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 1991;13(3):252–64.
41. Luan J, Zhang C, Xu B, Xue Y, Ren Y. The predictive performances of random forest models with limited sample size and different species traits. Fish Res 2020;227:105534.
42. Liebal UW, Phan ANT, Sudhakar M, Raman K, Blank LM. Machine learning applications for mass spectrometry-based metabolomics. Metabolites 2020;10(6):243.
43. Sathe SK, Seeram NP, Kshirsagar HH, Heber D, Lapsley KA. Fatty acid composition of California grown almonds. J Food Sci 2008;73(9):C607–14.
44. Subramanian C, Frank M, Batte J, Whaley S, Rock C. Oleate hydratase from *Staphylococcus aureus* protects against palmitoleic acid, the major antimicrobial fatty acid produced by mammalian skin. J Biol Chem 2019;294(23):9285–94.
45. Hagedoorn P, Hollmann F, Hanefeld U. Novel oleate hydratases and potential biotechnological applications. Appl Microbiol Biotechnol 2021;105(16–17):6159–72.
46. Feldman J, Lee E. Serotonin content of foods: effect on urinary excretion of 5-hydroxyindoleacetic acid. Am J Clin Nutr 1985;42(4):639–43.
47. Mashige F, Matsushima Y, Kanazawa H, Sakuma I, Takai N, Bessho F, et al. Acidic catecholamine metabolites and 5-hydroxyindoleacetic acid in urine: the influence of diet. Ann Clin Biochem 1996;33(1): 43–49.
48. Schmidt Andersen M-B, Kristensen M, Manach C, Pujos-Guillot E, Poulsen S, Larsen T, et al. Discovery and validation of urinary exposure markers for different plant foods by untargeted metabolomics. Anal Bioanal Chem 2014;406(7):1829–44.
49. Garcia-Aloy M, Llorach R, Urpi-Sarda M, Tulipani S, Estruch R, Martínez-González M, et al. Novel multimetabolite prediction of walnut consumption by a urinary biomarker model in a free-living population: the PREDIMED study. J Proteome Res 2014;13(7): 3476–83.
50. Tulipani S, Llorach R, Jáuregui O, López-Uriarte P, Garcia-Aloy M, Bullo M, et al. Metabolomics unveils urinary changes in subjects with metabolic syndrome following 12-week nut consumption. J Proteome Res 2011;10(11):5047–58.
51. Tohmola N, Johansson A, Sane T, Renkonen R, Hämäläinen E, Itkonen O. Transient elevation of serum 5-HIAA by dietary serotonin and distribution of 5-HIAA in serum protein fractions. Ann Clin Biochem 2015;52(4):428–33.
52. Ma S-R, Yu J-B, Fu J, Pan L-B, Yu H, Han P, et al. Determination and application of nineteen monoamines in the gut microbiota targeting phenylalanine, tryptophan, and glutamic acid metabolic pathways. Molecules 2021;26(5):1377.
53. Beloborodova N, Chernevskaya E, Getsina M. Indolic structure metabolites as potential biomarkers of non-infectious diseases. Curr Pharm Des 2021;27(2):238–49.
54. Fu Q, Tan Z, Shi L, Xun W. Resveratrol attenuates diquat-induced oxidative stress by regulating gut microbiota and metabolome characteristics in piglets. Front Microbiol 2021;12:695155.
55. Li N, Jia X, Chen C, Blumberg J, Song Y, Zhang W, et al. Almond consumption reduces oxidative DNA damage and lipid peroxidation in male smokers. J Nutr 2007;137(12):2717–22.
56. Hollis J, Mattes R. Effect of chronic consumption of almonds on body weight in healthy humans. Br J Nutr 2007;98(3):651–6.

57. Li S, Liu Y, Liu J, Chang W, Chen C, Chen C. Almond consumption improved glycemic control and lipid profiles in patients with type 2 diabetes mellitus. Metabolism 2011;60(4):474–9.
58. Jambazian P, Haddad E, Rajaram S, Tanzman J, Sabaté J. Almonds in the diet simultaneously improve plasma α-tocopherol concentrations and reduce plasma lipids. J Am Diet Assoc 2005;105(3):449–54.
59. USDA Agricultural Research Service (ARS). FoodData Central. Beltsville, MD: USDA ARS; 2019.
60. Stevens-Barrón JC, de la Rosa LA, Wall-Medrano A, Álvarez-Parrilla E, Rodríguez-Ramirez R, Robles-Zepeda RE, et al. Chemical composition and *in vitro* bioaccessibility of antioxidant phytochemicals from selected edible nuts. Nutrients 2019;11(10):2303.
61. Pearl J. Causal inference in statistics: an overview. Statist Surv 2009;3:96–146.
62. Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. New York: Cambridge University Press; 2015.
63. Frankenfeld CL, Hullar MAJ, Maskarinec G, Monroe KR, Shepherd JA, Franke AA, et al. The gut microbiome is associated with circulating dietary biomarkers of fruit and vegetable intake in a multiethnic cohort. J Acad Nutr Diet 2022;122(1):78–98.
64. Guasch-Ferré M, Hernández-Alonso P, Drouin-Chartier J-P, Ruiz-Canela M, Razquin C, Toledo E, et al. Walnut consumption, plasma metabolomics, and risk of type 2 diabetes and cardiovascular disease. J Nutr 2021;151(2):303–11.
65. Muller E, Algavi YM, Borenstein E. A meta-analysis study of the robustness and universality of gut microbiome-metabolome associations. Microbiome 2021;9(1):203.
66. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American Gut: an open platform for citizen science microbiome research. mSystems 2018;3(3):e00031–18.
67. Verdi S, Abbasian G, Bowyer RCE, Lachance G, Yarand D, Christofidou P, et al. TwinsUK: the UK Adult Twin Registry update. Twin Res Hum Genet 2019;22(6):523–9.
68. Park ST, Kim J. Trends in next-generation sequencing and a new era for whole genome sequencing. Int Neurourol J 2016;20(Suppl 2):S76–83.
69. Scalbert A, Brennan L, Manach C, Andres-Lacueva C, Dragsted LO, Draper J, et al. The food metabolome: a window over dietary exposure. Am J Clin Nutr 2014;99(6):1286–308.
70. Wishart D. FooDB version 1.0 University of Alberta, Edmonton, Canada. Date Accessed: 2022 Apr 14. [Internet]. Available from: https://foodb.ca/.
71. Rothwell JA, Perez-Jimenez J, Neveu V, Medina-Remón A, M'Hiri N, García-Lobato P, et al. Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. Database (Oxford) 2013: bat070.
72. PhytoHub. PhytoHub[Internet]. Available from: https://phytohub.eu/. Date Accessed: 2022 Apr 18.
73. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, et al. KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. Plant Cell Physiol 2012;53(2):e1.
74. Brennan L, Hu FB, Sun Q. Metabolomics meets nutritional epidemiology: harnessing the potential in metabolomics data. Metabolites 2021;11(10):709.