The Journal of Nutrition
Methodology and Mathematical Modeling

See corresponding commentary on page 273.

Fecal Bacteria as Biomarkers for Predicting Food Intake in Healthy Adults

Leila M Shinn,[1] Yutong Li,[2] Aditya Mansharamani,[3] Loretta S Auvil,[4] Michael E Welge,[4,5] Colleen Bushell,[4,5] Naiman A Khan,[1,6] Craig S Charron,[7] Janet A Novotny,[7] David J Baer,[7] Ruoqing Zhu,[2,4] and Hannah D Holscher[1,4,6,8]

[1]Division of Nutritional Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [2]Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [3]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [4]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [5]Mayo-Illinois Alliance for Technology-Based Healthcare, Urbana, IL, USA; [6]Department of Kinesiology & Community Health, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [7]Beltsville Human Nutrition Research Center, USDA Agricultural Research Service, Beltsville, MD, USA; and [8]Department of Food Science and Human Nutrition, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ABSTRACT

Background: Diet affects the human gastrointestinal microbiota. Blood and urine samples have been used to determine nutritional biomarkers. However, there is a dearth of knowledge on the utility of fecal biomarkers, including microbes, as biomarkers of food intake.

Objectives: This study aimed to identify a compact set of fecal microbial biomarkers of food intake with high predictive accuracy.

Methods: Data were aggregated from 5 controlled feeding studies in metabolically healthy adults ($n = 285$; 21–75 y; BMI 19–59 kg/m^2; 340 data observations) that studied the impact of specific foods (almonds, avocados, broccoli, walnuts, and whole-grain barley and whole-grain oats) on the human gastrointestinal microbiota. Fecal DNA was sequenced using 16S ribosomal RNA gene sequencing. Marginal screening was performed on all species-level taxa to examine the differences between the 6 foods and their respective controls. The top 20 species were selected and pooled together to predict study food consumption using a random forest model and out-of-bag estimation. The number of taxa was further decreased based on variable importance scores to determine the most compact, yet accurate feature set.

Results: Using the change in relative abundance of the 22 taxa remaining after feature selection, the overall model classification accuracy of all 6 foods was 70%. Collapsing barley and oats into 1 grains category increased the model accuracy to 77% with 23 unique taxa. Overall model accuracy was 85% using 15 unique taxa when classifying almonds (76% accurate), avocados (88% accurate), walnuts (72% accurate), and whole grains (96% accurate). Additional statistical validation was conducted to confirm that the model was predictive of specific food intake and not the studies themselves.

Conclusions: Food consumption by healthy adults can be predicted using fecal bacteria as biomarkers. The fecal microbiota may provide useful fidelity measures to ascertain nutrition study compliance. *J Nutr* 2021;151:423–433.

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

Introduction

Research is increasingly demonstrating that the gastrointestinal microbiota affects human health (1, 2). The gastrointestinal microbiota is influenced by genetic, physiological, and environmental factors (3). Diet is a modifiable environmental factor that affects the composition of the gastrointestinal microbiota (4, 5). Fruits, vegetables, whole grains, and nuts are incompletely digested and, thus, serve as substrates for microbial metabolism (3). Indeed, the measured metabolizable energy of almonds is 25% less than the estimated value using Atwater specific factors (6), the metabolizable energy from walnuts is 21% less than that predicted (7), and a grain-based diet provides ~8% less energy than predicted (8). We previously reported that consumption of walnuts (9), almonds (10), and avocado (11) affected bacterial genera within the Firmicutes phylum, whereas broccoli (12) consumption increased the relative abundance of genera in the Bacteroidetes phylum. Research examining barley consumption revealed an increase in relative abundances of Clostridiaceae, *Roseburia*, and *Ruminococcus* (13). This growing body of research indicates that foods that contain nondigestible nutrients and fiber differentially affect the human gastrointestinal microbiota.

Manuscript received April 1, 2020. Initial review completed June 8, 2020. Revision accepted August 27, 2020.
First published online October 6, 2020; doi: https://doi.org/10.1093/jn/nxaa285.

Self-reported measures of food intake and compliance are frequently utilized in nutrition research. Their benefits include the relative ease of collection and low costs (14, 15); however, there are challenges related to these self-reported measures that limit their ability to accurately measure food intake (16). Therefore, objective biomarkers that can complement or replace self-reported measures of food intake are of interest. Biomarkers for nutrient intake, like urinary sodium (sodium intake), nitrogen (protein intake), and energy, have been a major focus in the field (17, 18). Previous work has demonstrated the efficacy of various nutrients as food-specific biomarkers, including lutein, carotenoids, tocopherols, folate, vitamin B-12, and phospholipid fatty acids (19–21). Whereas there are reports on the use of urinary and blood biomarkers for specific foods and food groups (22–26), there is a dearth in knowledge on the utility of fecal samples, a noninvasive biological sample, to generate nutritional biomarkers. Importantly, the ability to easily collect fecal samples on a population level was demonstrated by the American Gut Project, which has collected >11,000 fecal samples from US citizens across the lifespan (27). However, to our knowledge, bacterial biomarkers of food intake have not been reported.

Ultimately, the unique impact of whole foods on the human gastrointestinal microbiota could allow for the development of noninvasive microbial biomarkers of food intake. Thus, we aimed to develop a proof-of-concept predictive model to identify fecal bacterial biomarkers of intake of specific foods that had previously been shown to affect the human gastrointestinal microbiota, including almonds, avocados, broccoli, walnuts, and whole grains. Herein, we describe analyses conducted on data generated from fecal samples collected at the baseline and end of 5 feeding trials. The purpose of the present investigation was to utilize a computationally intensive, multivariate, machine learning approach to identify bacterial biomarkers with high predictive accuracy of food intake.

Methods

Participants and treatments

Data were from 5 separate feeding studies examining almond, avocado, broccoli, walnut, or whole-grain barley and whole-grain oat consumption in adults (n = 285) between 21 and 75 y of age (**Supplemental Table 1**). 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).

Almond.

Methods undertaken in the almond trial (NCT02034383) were previously described (6). Briefly, 18 metabolically healthy adults [25–75 y, BMI (in kg/m^2): 21.9–36.1] underwent a completely controlled-feeding, randomized, crossover trial with five 3-wk periods separated by 1-wk washouts between each period. Five treatments were completed including *1*) 0 servings/d of almonds (control), *2*) 1.5 servings (42 g)/d of whole almonds; *3*) 1.5 servings/d of whole, roasted almonds; *4*) 1.5 servings/d of roasted, chopped almonds; and *5*) 1.5 servings/d of almond butter (not included in these analyses). The almonds provided 257 kcal and 15% of the percentage Daily Value for dietary fiber (28). Fecal samples were collected before the intervention and at the end of each diet period.

Avocado.

Metabolically healthy adults (n = 163; 25–45 y, BMI: 23.9–58.8) underwent an investigator-blinded, parallel-arm, randomized, controlled trial (NCT02740439), as previously described (11, 29). Participants consumed isocaloric meals with or without avocado once daily for 12 wk. Men in the avocado group consumed 175 g avocado/d [192 kcal; 42% of the DV for dietary fiber (28)], whereas women consumed 140 g/d [153 kcal; 33% of the DV for dietary fiber (28)]. Fecal samples were collected at baseline (before the intervention) and at the end of the 12-wk intervention period.

Broccoli.

Methods undertaken for this trial (NCT02346812) were previously described (30). Briefly, 18 metabolically healthy adults (21–70 y, BMI: 19.0–36.6) underwent a completely controlled-feeding, randomized, crossover trial consisting of two 18-d treatment periods separated by a 24-d washout. The 2 treatments were *1*) a Brassica-free control diet and *2*) the same diet with 200 g of cooked broccoli and 20 g of fresh daikon radish per day. Broccoli provided 68 kcal and 19% of the DV for dietary fiber (28). Fecal samples were collected before the intervention and at the end of each treatment period.

Walnut.

Methods undertaken for this trial (NCT01832909) were previously described (7). Briefly, 18 metabolically healthy adults (25–75 y, BMI: 20.2–34.9) underwent a completely controlled-feeding, randomized crossover trial. Two isocaloric diet periods were completed including *1*) 0 g walnuts/d (control) and *2*) 42 g walnuts/d for two 3-wk periods, with a 1-wk washout between diet periods. The walnuts provided 275 kcal and 10% of the DV for dietary fiber (28). Fecal samples were collected before the intervention of this controlled feeding trial and at the end of each diet period.

Whole-grain oats and barley.

Metabolically healthy adults (n = 68; 25–70 y, BMI: 18.9–38.3) underwent a 6-wk randomized, double-blinded, parallel-arm, completely controlled-feeding trial (NCT01293604) (31). The isocaloric conditions were *1*) 0.7 daily servings (11.2 g) of whole grain/1800 kcal (control), *2*) 4 daily servings (64 g) of whole-grain barley/1800 kcal, or *3*) 4 daily servings (64 g) of whole-grain oats/1800 kcal. The barley treatments provided 227 kcal and 40% of the DV for dietary fiber (28). The oats provided 243 kcal and 23% of the DV for dietary fiber (28). Fecal samples were collected before the intervention and at the end of each diet period.

Microbiota composition

Fecal samples were collected at the beginning and end of each dietary period for all 5 studies. Fecal sample collection, DNA extraction, and 16S ribosomal RNA (rRNA) gene (V4 region) sequencing were conducted as previously described (11). There was a mean ± SD of 50,053 ± 37,433 sequences per sample (range: 3764–373,284). Sequences from all 5 studies were analyzed together with bioinformatics software [Divisive Amplicon Denoising Algorithm 2 (DADA2) and Quantitative Insights Into Microbial Ecology 2 (QIIME 2, version 2018.6)] (32, 33). Briefly, DADA2 was used to denoise and dereplicate

Supported by Foundation for Food and Agriculture Research New Innovator Award (to HDH), USDA National Institute of Food and Agriculture Hatch Project 1009249 (to HDH), an ACES Jonathan Baldwin Turner Fellowship (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: The authors report no conflicts of interest.
Supplemental Tables 1 and 2 and Supplemental Figures 1–4 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/.
The entire R script of this study and further statements on the data can be accessed from Github at "*cralo31/foodML*."
LMS and YL contributed equally to this work.
Address correspondence to HDH (e-mail: hholsche@illinois.edu) or RZ (e-mail: rqzhu@illinois.edu).
Abbreviations used: DADA2, Divisive Amplicon Denoising Algorithm 2; DV, Daily Value; PC, principal component; rRNA, ribosomal RNA.

the sequences, trim off the primers, and remove chimeras and reads with a quality score <20. A total of 4375 amplicon sequence variants were identified. Taxonomy was assigned using SILVA release 132 (34).

The raw taxonomic data set contained 742 observations, including the baseline and end data for all subjects across the 5 feeding studies (**Supplemental Figure 1**). First, avocado subjects with missing baseline or end sequence data were removed (718 observations remaining). Next, the sequence data from the almond butter group in the almond trial were removed owing to the Atwater values revealing the butter was completely digested (35) and there being no differences between the control and almond butter fecal microbiota (10), resulting in 682 remaining observations. Two additional almond observations were dropped owing to missing baseline or end data for the roasted and chopped treatment arms, respectively (680 observations remaining). Finally, the differences between baseline and end relative abundances for those 680 observations were calculated to generate 340 observations for subsequent analyses. These 340 observations contain the treatment data set ($n = 199$), i.e., the study periods where study diets contained the specific foods, and the control data set ($n = 141$).

Statistics

Feature selection and classification.

The predeclared primary endpoint, bacterial species as biomarkers of food intake, was not changed during the course of this study. First, we considered the changes in relative abundance (treatment end minus baseline) for each of the features (fecal bacterial species) to perform the analyses. Using the difference between baseline and postintervention also allows for determination of the internal differences or batch effects related to performance of these studies at 2 different sites and the background diet, as well as reducing biases related to utilization of only the study endpoints (36). Furthermore, analyzing the difference in abundances results in a normal distribution centered at 0, making it more suitable to model.

In addition to investigating whether an individual's food intake can be predicted from changes in fecal microbial composition, a secondary aim of the study was to identify a compact set of bacteria that are mainly driving the prediction of the diet for interpretation purposes. Thus, we conducted feature selection to create a smaller feature space with high classification performance. This was done by first reducing the number of features (fecal bacterial species) by removing sparse features (species) with predominantly zero values. Removal of the sparse features also prevented the random forest model from simply considering the nonzero values compared with the zero values, which would bias the classification, and increase the noise and computational inefficiency of the analyses.

Next, we performed marginal screening using the Kruskal–Wallis test (37) using each species as the covariate (predictor) and the specific food (i.e., almonds, avocados, broccoli, walnuts, whole-grain barley, or whole-grain oats) as the outcome, which allowed for assessment of the difference in distribution between each control (absence of study-specific food) and treatment group (inclusion of study-specific food). The 20 most significant bacterial taxa differentiating the treatment groups from their corresponding control groups were selected from each of the 5 studies, totaling 89 distinct features. According to our numerical experiments, choosing 20 significant features per food was sufficient to capture the main signal of the microbiota, with no evident improvement in performances if more features were considered.

Next, we used random forests to model these data and further select the set of features that distinguished the treatment labels. This was done with the *randomForest* package (38). For tuning the random forest, the number of trees was set to 2000 and the node sizes range from 1 to 10. The "mtry" parameter was set to the default value of $\sqrt{p}$, where p is the number of features considered for the model. We then implemented grid-search across the considered tuning parameters to choose the best model using the out-of-bag estimates (39) generated from the random forest model. The results and variable importance from this step revealed that not all of the features (20 features from each food) identified during marginal screening were predictive in the random forest model. Thus, the number of species used as features was further decreased using variable importance scores generated from the initial random forest to determine a compact feature set with a minimum loss in classification accuracy. The top 20, 15, 12, 10, 7, 5, 3, and 2 species (as ranked by variable importance) from each study were selected and pooled. The number of resulting unique species varied according to the number of foods considered. **Supplemental Figure 2** provides further details of this model selection approach to determining the final number of features-per-food (and consequently, the number of unique features overall).

Selecting 10 species from each of the 6 foods and pooling resulted in 22 unique bacterial species. Collapsing barley and oats into a grains category (5 foods) resulted in 23 species after feature selection. Removing broccoli from the data set (4 foods) reduced the number of unique species remaining after feature selection to 15. The unique species selected (either 22, 23, or 15) from pooling the top 10 most significant species for each food were refit into a second random forest using the same parameters to classify the specific foods consumed.

Removing the study-specific batch effect.

To ensure that the results signified the food consumed rather than participation in a specific study, we quantified and removed the batch effect. We consider a shift in microbiota composition in the treatment groups (i.e., the treatment signal) to be driven by 2 separate effects. The first is the effect of the food (e.g., almond), which we call the treatment effect. The second is the effect of the background diet (the batch effect), which is indicative of all the foods each participant consumed aside from the study-specific food of interest. We considered any bacterial signatures present in both control and treatment subjects as a batch effect, because their presence could artificially inflate the classification accuracy of the models. Because the objective was to predict intake of the 6 foods (i.e., almonds, avocados, broccoli, walnuts, whole-grain barley, and whole-grain oats), the food effect signal was extracted from the entirety of the treatment signal. We removed the batch effect from the feature set by removing the influence of the first few principal components (PCs) of the control group data from the treatment group data (40). The precise number of PCs removed from each of the food groups (6, 5, and 4 foods) was determined via the elbow method (41) (**Supplemental Figure 3**).

Next, we performed singular value decomposition on the feature set for all participants in the control group from each food:

$$X_{\text{cont}} = U_{\text{cont}}\, D_{\text{cont}} V_{\text{cont}}^{T} \tag{1}$$

Then, the treatment group feature set was projected onto the column space spanned by the PCs of the control group by multiplying the projection matrix:

$$\tilde{X}_{\text{treat}} = X_{\text{treat}} V_{\text{cont}} V_{\text{cont}}^{T} \tag{2}$$

Finally, $\tilde{X}_{\text{treat}}$ was subtracted from X_{treat}, effectively removing the main signal within the control group from the treatment group. If the study foods (and not just the studies) altered the composition of the microbiota in different directions, we expected the classification performance to be similar when using treatment or treatment effect. **Figure 1** provides a theoretical demonstration of this batch effect.

Additional calculations were conducted to validate the food prediction approach. The random forest model built previously was used to classify the control subjects from each trial to predict the food consumed, using the same methodology applied to treatment data. These classification results were then compared with baseline performance as an approach to ensure that the model was predicting food consumption rather than trial participation.

Results

Because the objective of this study was to maximize classification accuracy with a minimal set of features, or biomarkers, from each food (i.e., almonds, avocados, broccoli, walnuts, and whole-grain barley and whole-grain oats), we first examined the influence of the number of features selected on model accuracy.

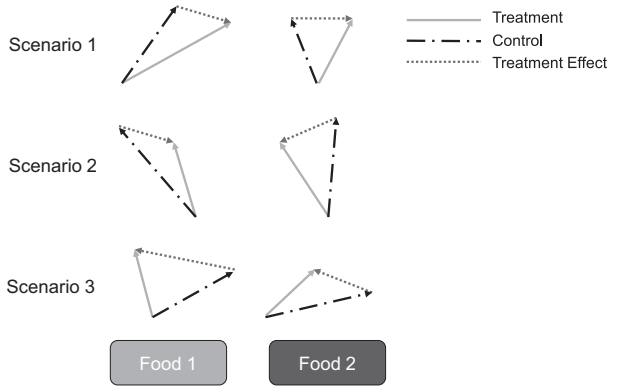

FIGURE 1 Theoretical representation of the batch effect. The batch effect occurs when the control and treatment groups exhibit different changes across different study sites. The measured signal within the control group represents the background conditions of the study, and the measured treatment signal is comprised of the background conditions of the study and the effect of the specific food. Scenario 1 depicts the case where all of the directions are the same for both foods. The measured treatment signal (solid line) is similar to the treatment effect, or true effect, of the food (dashed line). In this scenario, controlling for the control, or batch effect, will not have an impact on the classification accuracy. Scenario 2 represents the case where treatments from both foods are pointing in the same direction. However, the treatment effects between the 2 are different from one another because the control groups move in different directions. If we correct for the control group in scenario 2, we are able to account for the true treatment effect and correctly separate the control from the treatment. However, if we do not adjust for the background diet effect here, we will not observe the effect, leading to a false negative (type II error). Finally, in scenario 3, both treatment groups are moving in different directions from baseline, but the treatment effects of the 2 foods are in the same direction. Therefore, if the background diet effect is not accounted for in this scenario, we observe differences between the 2 groups, when in reality, there is not one (type I error).

Selecting more than the 10 important features from each study group for model training did not provide a significant increase in classification performance, i.e., the classification accuracy was 71% when using 10 features compared with 69% when using 20 features for each of the 6 foods (Supplemental Figure 2).

Furthermore, the *P* values from the paired-difference cross-validation *t* test between the models with >10 features were not significant. Because the data set is unbalanced (**Supplemental Figure 4**)—the avocado study comprises the largest number of participants and datapoints—we also considered the AUC as a performance criterion for classification.

Table 1 reports the top 23 bacterial species (selected from the highest 10 variable importance scores from the random forest model for each study) that were used in the analyses. Species assignments are listed in descending rank order as determined via variable importance from the random forest model.

Tables 2–4 are formatted as follows. *With batch effect* rows contain the confusion matrices, per-class accuracy, and overall accuracy of a random forest model trained on the differences in relative abundance of the taxa selected as discussed in the Statistics section. *Removed batch effect* rows contain the confusion matrices, per-class accuracy, and overall accuracy of the same random forest model trained on the same data set with the batch effect removed. Finally, *Control* rows contain the confusion matrices, per-class accuracy, and overall accuracy of the same random forest model when predicting the foods using the control subjects rather than the participants that consumed the study foods, i.e., the participants in the treatment conditions of each study. Because the control subjects should exhibit none of the signal learned by the training data, the results reported in the control rows of each table indicate the baseline performance of that machine learning model. Effectively, baseline performance indicates how well a machine learning model could classify inputs given useless (or no) training data. The classification accuracy of the models in the control rows (near 50%) corresponds to the proportion of avocado training samples (near 50%) (Supplemental Figure 4). Because avocado is the largest group within the training data, we expected that this model would classify almost all control subjects into the avocado group, confirming the training data had the batch effect removed.

The performance seen in the *With batch effect* and *Removed batch effect* rows of Tables 2–4 is then compared with the baseline performance of the control group (*Control* rows) to determine if the machine learning models were given useful training data. If the *Control* rows are similar to the other 2 models (*With batch effect* and *Removed batch effect* rows), it would indicate that the models had learned no useful information from the training data, i.e., how to predict which specific food was eaten. Ideally, the performance seen in the *With batch effect* and *Removed batch effect* rows should largely exceed the baseline performance metric in the *Control* rows. In addition to the prediction accuracy, we also present multiclass classification AUC results using the *pROC* package (42).

With all 6 foods in the model, the data set had adequate signal to predict the foods consumed, resulting in an overall model accuracy of 70%, with broccoli being confused for avocado 89% of the time and oats and barley frequently being confused for one another as shown in Table 2 (*With batch effect*). After removing the batch effect, we re-examined classification performance. If classification performance were to fall drastically, that would indicate the original classification accuracy using just the data from participants in the treatment condition was artificially inflated and the model was only performant at detecting which research study a given subject participated in. Overall model accuracy fell to 66% as shown in Table 2 (*Removed batch effect*). However, the reduction in classification accuracy from the initial models (Table 2, *With batch effect*) to the models without the batch effect present (Table 2, *Removed batch effect*) was not large. Because classification accuracy did not fall drastically when the batch effect was removed from the training data, we concluded that the effect of consuming a specific food exhibited an adequate bacterial signal in each study participant, leading to ideal classification performance.

In addition to retaining classification accuracy after removal of the batch effect, we were able to further confirm the validity of the results by comparing the classification accuracy of control subjects with baseline results. In this case, baseline results correspond to classification of the specific food, given there was no signal separating the foods within the data set (i.e., it would be a random guess as to which food was consumed). To maximize classification accuracy given no data, the best method is to classify every subject as belonging to the largest group present in the training data, which was the avocado group in the current report. To confirm that the models were trained on data with a minimal batch effect, we compared the results of classifying control subjects with the baseline results as shown in Table 2 (*Control*). As expected, the model classified almost all of the subjects into the avocado group using control data, because

TABLE 1 Bacterial biomarkers using the top fecal microbiota species from metabolically healthy adults who consumed 4 foods, 5 foods, and 6 foods[1]

	Rank (overall variable importance)[2]		
SILVA assignment	4 Foods	5 Foods	6 Foods
Roseburia undefined	1 (0.097)	1 (0.077)	1 (0.069)
Lachnospira spp.	2 (0.043)	3 (0.034)	5 (0.021)
Oscillibacter undefined	3 (0.039)	4 (0.032)	4 (0.023)
Subdoligranulum spp.	4 (0.039)	5 (0.030)	3 (0.025)
Streptococcus salivarius subsp. *thermophilus*	5 (0.039)	2 (0.035)	2 (0.026)
Parabacteroides distasonis	6 (0.032)	7 (0.023)	9 (0.015)
Roseburia spp.	7 (0.026)	6 (0.025)	7 (0.017)
Anaerostipes spp.	8 (0.023)	8 (0.019)	6 (0.020)
Lachnospiraceae *ND3007* group undefined	9 (0.022)	9 (0.018)	8 (0.016)
Ruminiclostridium spp.	10 (0.022)	10 (0.015)	10 (0.011)
Lachnoclostridium undefined	11 (0.013)	13 (0.009)	15 (0.005)
Undefined species in Clostridiales order	12 (0.011)	12 (0.010)	—
Faecalibacterium undefined	13 (0.010)	14 (0.008)	12 (0.007)
Ruminococcaceae *UCG-013* undefined	14 (0.008)	15 (0.006)	13 (0.006)
Lachnospiraceae *UCG-001* undefined	15 (0.006)	18 (0.004)	21 (0.001)
Bacteroides spp.	—	11 (0.013)	14 (0.005)
Ruminiclostridium spp.	—	16 (0.005)	11 (0.007)
Butyricimonas undefined	—	17 (0.004)	17 (0.003)
Dialister spp.	—	19 (0.003)	16 (0.004)
Ruminococcaceae *NK4A214* spp.	—	20 (0.002)	18 (0.003)
Hydrogenoanaerobacterium spp.	—	21 (0.002)	20 (0.002)
Lachnospiraceae *UCG-001* spp.[3]	—	22 (0.001)	—
Intestinimonas spp.	—	23 (0.001)	22 (0.001)
Sutterella spp.[4]	—	—	19 (0.003)

[1] Top 10 species-level SILVA taxa from each of the 4 foods (collapsed grains category, broccoli removed; 15 total taxa); 5 foods (collapsed grains category; 23 total taxa); and 6 foods (22 total taxa).
[2] Variable importance scores were generated from each food using the initial random forest model to determine a compact feature set with a minimum loss in classification accuracy. The top 20, 15, 12, 10, 7, 5, 3, and 2 species (as ranked by variable importance) from each study were selected and pooled. The number of resulting unique species varied. Selecting >10 species from each study did not significantly increase classification accuracy.
[3] Unique to 5 foods analyses.
[4] Unique to 6 foods analyses.

the avocado trial had the largest sample size. Classification of the control group yielded an overall baseline accuracy of 50% as shown in Table 2 (*Control*).

These same interpretations can be applied to the results shown in Tables 3 and 4. Aggregating oats and barley into 1 category, whole grains, improved the classification accuracy of the whole grain category to 94% and improved the overall model accuracy to 77%, as shown in Table 3 (*With batch effect*). Overall model accuracy after removing the batch effect was 73%, and the baseline accuracy was 55% as shown in Table 3 (*Removed batch effect* and *Control*, respectively).

Removing broccoli from the model due to low classification accuracy, resulting in 4 foods, increased the model accuracy to 85% as shown in Table 4 (*With batch effect*). Removing the batch effect reduced this model accuracy to 76% as seen in Table 4 (*Removed batch effect*), and the baseline accuracy was 55% as seen in Table 4 (*Control*).

Discussion

Although diet–microbiota studies have demonstrated that specific foods affect the fecal microbiota, to our knowledge, none have used machine learning to generate bacterial biomarkers that predict specific food consumption. Because the composition of the gastrointestinal microbiota is complex, machine learning models have become a novel tool for extracting useful information from these large data sets (43). This study predicted dietary intake of specific whole foods (i.e., almonds, avocados, broccoli, walnuts, whole-grain barley, and whole-grain oats) with 70%–85% accuracy using 22 and 15 fecal bacteria as biomarkers of food intake, respectively. The maximal model accuracy, 85% accuracy, was achieved when whole-grain oats and whole-grain barley were collapsed into 1 whole grains category and broccoli was removed from the model (4 food groups). Examining the classification accuracy after removing the batch effect and the baseline model, these results revealed the ability to predict the food consumed, rather than research trial participation. From these results, we can conclude that the model did not falsely detect the bacterial signatures of the foods within the control group. Conversely, because the same model performed well on the treatment groups, we can also conclude it was trained using data with adequate signal strength to determine the effects of the studied foods.

Biomarkers for nutrient intake, like urinary sodium (sodium intake), nitrogen (protein intake), and energy, have been a major focus in the nutrition field (17, 18). Previous work has demonstrated the efficacy of various nutrients serving as food-specific biomarkers, including lutein as a biomarker of avocado intake (29) and tocopherols as a biomarker for almonds intake (20). Although urinary and blood biomarkers have been used for specific foods and food groups (22–26), there is a dearth

TABLE 2 Prediction of specific food intake from metabolically healthy adults who consumed 6 foods using random forest with batch effect, removal of batch effect, and control groups[1]

	Almond, *n*	Avocado, *n*	Broccoli, *n*	Walnut, *n*	Barley, *n*	Oats, *n*	Accuracy, %
Almond							
With batch effect[2]	37	12	0	0	0	0	76
Removed batch effect[3]	38	11	0	0	0	0	78
Control[4]	1	13	0	0	4	4	6
Avocado							
With batch effect[2]	5	59	2	0	1	0	88
Removed batch effect[3]	10	55	1	0	0	1	82
Control[4]	2	62	0	0	2	2	94
Broccoli							
With batch effect[2]	0	16	1	0	1	0	6
Removed batch effect[3]	2	12	4	0	0	0	22
Control[4]	0	16	0	0	2	2	0
Walnut							
With batch effect[2]	1	2	0	13	2	0	72
Removed batch effect[3]	3	3	0	10	2	0	56
Control[4]	1	10	0	0	7	7	0
Barley							
With batch effect[2]	0	1	0	0	18	5	75
Removed batch effect[3]	0	3	0	1	13	7	54
Control[4,5]	1	3	0	0	17	17	81
Oats							
With batch effect[2]	0	2	0	0	10	11	48
Removed batch effect[3]	1	4	0	1	6	11	48
Control[4]	1	3	0	0	17	17	81
Overall accuracy and AUC							
With batch effect[2]							70 (AUC: 0.92)
Removed batch effect[3]							66 (AUC: 0.89)
Control[4]							57 (AUC: 0.78)

[1] All analyses utilized the top 10 species-level SILVA taxa from each of the 6 foods (22 total taxa).
[2] Values from random forest classification model with batch effect present predicting 6 foods.
[3] Values from random forest classification model removing batch effect (via removal of principal components) predicting 6 foods.
[4] Values from random forest classification model predicting 6 foods from control group participants.
[5] Whole-grain barley and whole-grain oats control groups were the same.

in knowledge on the utility of fecal samples, a noninvasive biological sample, to generate nutritional biomarkers. Ultimately, the unique impact of whole foods on the human fecal microbiota, metagenome, and metabolome could allow for the development of noninvasive microbial biomarkers of food intake to complement current practices.

Although, to our knowledge, bacterial biomarkers have not previously been utilized to predict food intake, others have utilized fecal microbiota as biomarkers for health status, including glycemic responses (44), and variability in BMI, triglyceride, and HDL cholesterol concentration (45). In addition, microbes in the oral cavity, gastrointestinal tract, and pancreatic tissue differed in patients with pancreatic cancer compared with healthy controls (46). Also, Das et al. (47) linked microbial metabolites to inflammatory bowel disease and Turnbaugh et al. (48) revealed changes in fecal microbial structure, gene expression, and subsequent metabolic pathways due to dietary changes. Although these studies are not without limitations, they reveal promise in the utility of fecal samples to fill the need for noninvasive, inexpensive, and specific markers of dietary intake and health status (49). Importantly, our results help fill the gap in knowledge related to the utility of fecal microbial biomarkers for food intake.

Random forest was performant in the classification of various labels based on microbial composition of the microbiome (50). Theoretically, it may be possible to accomplish the same goal using traditional linear models by adding the batch effect as a covariate (50, 51); however, linear modeling does not adequately account for multicategorical outcomes like the random forest (52). In addition, we also considered the lasso model (53) initially for our study, but the results were inferior to the random forest (**Supplemental Table 2**). Random forest models can also model highly nonlinear relations leading to better classification performance (54). In addition to their robustness for classification, random forest models provide methods for feature selection by assigning variable importance scores to each input feature (e.g., the change in relative abundances for specific taxa across all subjects), allowing for an easy reduction of the feature space by eliminating the least important variables. Conversely, the use of random forest necessitated additional methods to remove the batch effect and use cross-validation to assess false detections.

The highest model performance was observed after aggregating whole-grain oats and barley into a single category (whole grains) and removing broccoli, which was often misclassified as avocado. Given that oats and barley contain similar insoluble and soluble fibers (mainly β-glucans), resistant starch, lipids, proteins, and phenolic content, we ultimately aggregated the 2 into 1 whole grains category (55, 56). Broccoli misclassification may be due to a number of factors, such as small sample size ($n = 18$) or similar nutrient composition to the other foods investigated, thereby resulting in an inadequate bacterial

TABLE 3 Prediction of specific food intake from metabolically healthy adults who consumed 5 foods using random forest with batch effect, removal of batch effect, and control groups[1]

	Almond, *n*	Avocado, *n*	Broccoli, *n*	Walnut, *n*	Grains, *n*	Accuracy, %
Almond						
With batch effect[2]	37	12	0	0	0	76
Removed batch effect[3]	38	11	0	0	0	78
Control[4]	1	10	0	0	7	6
Avocado						
With batch effect[2]	6	57	3	0	1	85
Removed batch effect[3]	9	55	0	0	3	82
Control[4]	2	51	0	0	13	77
Broccoli						
With batch effect[2]	0	15	2	0	1	11
Removed batch effect[3]	2	13	1	0	2	6
Control[4]	0	13	0	0	5	0
Walnut						
With batch effect[2]	1	2	0	13	2	72
Removed batch effect[3]	3	4	0	9	2	50
Control[4]	0	4	0	0	14	0
Grains						
With batch effect[2]	0	3	0	0	44	94
Removed batch effect[3]	0	5	0	0	42	89
Control[4,5]	0	2	0	1	18	86
Overall accuracy and AUC						
With batch effect[2]						77 (AUC: 0.93)
Removed batch effect[3]						73 (AUC: 0.90)
Control[4]						50 (AUC: 0.77)

[1] Barley and oats were collapsed into 1 "grains" category. All analyses utilized the top 10 species-level SILVA taxa from each of the 5 foods (23 total).
[2] Values from random forest classification model with batch effect present predicting 5 foods.
[3] Values from random forest classification model removing batch effect (via removal of principal components) predicting 5 foods.
[4] Values from random forest classification model predicting 5 foods from control group participants.
[5] Whole-grain barley and whole-grain oats control groups were the same.

signature to allow for accurate food intake classification. Specific to the similarities in nutrient composition contributing to the incorrect classification of broccoli consumption as avocado consumption, both avocados and broccoli are sources of pectin, lutein, and β-sitosterol (57–65). The inability to accurately classify broccoli intake from 16S rRNA bacterial sequence data may also be due to phenotypic responses (66). Future work is necessary to better understand the individualized nature of diet–microbiota interactions.

The high classification accuracy of the model points to the impact of specific whole foods on the intestinal microbiota, which is largely due to their supply of nondigested nutrients and fiber that is available for microbial metabolism. Specific to our work, almonds, broccoli, and walnuts are good sources of dietary fiber (providing 10%–19% of the recommended daily intake per reference amount customarily consumed), whereas avocado, whole-grain oats, and whole-grain barley are excellent sources of dietary fiber (providing ≥20% of the recommended daily intake per reference amount customarily consumed) (28). On a daily basis, ~40 g of dietary carbohydrates, 12–18 g of protein, and 5% of dietary lipids are not digested and reach the colon (67). Because the enzymatic capacity of the gastrointestinal microbiome far exceeds that of the human genome, these nondigested components are able to serve as substrates for a range of microbes, resulting in a signature of fecal microbial changes due to dietary intake. For example, certain PUFAs, which are found in almonds, avocados, walnuts, and whole grains, can be metabolized by intestinal microbes. In vitro studies examining *Roseburia* spp. revealed their ability to conjugate linoleic acid (68, 69). Furthermore, the intestinal microbiota partially controls the synthesis and regulation of bile acid secretion (70). The type of dietary fat also affects the interactions between bile acids and the intestinal microbiota (71). For example, walnut consumption resulted in reductions in the microbially derived secondary bile acids, deoxycholic acid and lithocholic acid (9). Participants who consumed avocado had lower concentrations of the bile acids, cholic and chenodeoxycholic acid, than controls (11). These unique host–microbe interactions, which are facilitated by the nutrients in specific foods, contribute to the signatures of food consumption and ability to create a panel of fecal bacterial biomarkers of food intake.

The bacterial species identified as the most relevant biomarkers for predicting food consumption align with previous findings. Specifically, the relative abundance of *Roseburia* spp. was significantly increased in previous studies examining the impact of almond, walnut, and whole-grain barley and oat intake on the microbiota (9, 10, 13, 72, 73). In addition, the relative abundance of *Lachnospira* spp. was increased with almond and avocado intake and identified as a predictive feature in the current report (10, 11). Furthermore, participants who consumed avocado had enriched *Faecalibacterium* (11). Others have also reported increased abundances of *Lachnospira* spp. after whole grain (73) and whole-grain barley consumption (13). Also, walnuts were reported to significantly increase the abundance of the Ruminococcaceae family when compared with a nut-free control (74). It is important to note that some of these previous findings utilize the same data as the current effort (9–11). Therefore, although the sequence data were reanalyzed in the present report, the similarity in findings

TABLE 4 Prediction of specific food intake from metabolically healthy adults who consumed 4 foods using random forest with batch effect, removal of batch effect, and control groups[1]

	Almond, *n*	Avocado, *n*	Walnut, *n*	Grains, *n*	Accuracy, %
Almond					
With batch effect[2]	37	12	0	0	76
Removed batch effect[3]	35	12	1	1	71
Control[4]	2	10	0	6	11
Avocado					
With batch effect[2]	5	59	0	3	88
Removed batch effect[3]	3	57	1	6	85
Control[4]	1	54	0	11	82
Walnut					
With batch effect[2]	1	2	13	2	72
Removed batch effect[3]	5	3	9	1	50
Control[4]	0	10	0	8	0
Grains					
With batch effect[2]	0	2	0	45	96
Removed batch effect[3]	1	9	0	37	79
Control[4]	0	8	1	12	57
Overall accuracy and AUC					
With batch effect[2]					85 (AUC: 0.95)
Removed batch effect[3]					76 (AUC: 0.89)
Control[4]					55 (AUC: 0.69)

[1] Barley and oats were collapsed into 1 "grains" category and broccoli was removed from the model. All analyses utilized the top 10 species-level SILVA taxa from each of the 4 foods (15 total).
[2] Values from random forest classification model with batch effect present predicting 4 foods.
[3] Values from random forest classification model removing batch effect (via removal of principal components) predicting 4 foods.
[4] Values from random forest classification model predicting 4 foods from control group participants.

is expected. Previous research demonstrating the importance of these microbes in relation to the consumption of the specific foods within our current effort shows the promise in future research utilizing these and other microbes as biomarkers of food intake.

The major strength of this work is that 4 of the 5 studies from which data were aggregated were randomized, controlled, crossover studies that provided participants with all of the foods consumed during the study period. Therefore, the only difference between the treatment and control diets was the addition of the studied food, i.e., almonds, broccoli, walnuts, and whole-grain barley or whole-grain oats. This is the strongest possible design for studying the impact of consumption of specific foods on the human gastrointestinal microbiota. Although the controlled feeding component of the trials included in this study is a major strength, it may limit translation of these findings to studies of the general population owing to selection biases and external validity issues inherent to randomized controlled trials. Participants in the current study were metabolically healthy adults 21–75 y of age with a wide BMI range (19–59). Therefore, although our sample encompasses a wide range of ages and BMI categories, findings cannot be generalized to children or those with various disease states. As such, future studies that include more diverse populations are needed for external validity. It is also important to note that our model does not reveal a quantitative measure of intake, but rather a qualitative measure. In addition, our approach utilized a change in relative abundance of taxa in our treatment groups compared with their control groups to identify bacteria of interest, which may pose limitations in applying these methods to observational studies. However, others have utilized semiquantitative methods that have been deemed appropriate for exploratory studies aimed at identifying biomarkers of food intake to examine their robustness in observational trials (23). More human intervention studies with varying doses of specific foods and quantitative methods of analysis are necessary to develop complete biomarkers. Future research should re-examine the impact of the foods examined herein in independent cohorts, in addition to new foods of interest, to validate the reliability and generalizability of our classification model. Furthermore, cooking and processing alter the physiochemical properties of food, which can result in differential impact on the gastrointestinal microbiota (10, 75). Therefore, food form should be considered in future studies.

Potential future work could consider different machine learning models to address the dietary predictive question (or similar issues). The main reason we utilized random forest is because of its ease of usage for addressing multiclass classification problems. This problem is a challenge in and of itself, which narrows the selection of choices for viable models. Nevertheless, methods such as boosting (76, 77) or angle-based support vector machines (78) are possible alternatives and may be worth exploring in future studies. Deep learning (79) also handles such unbalanced multiclass classification problems well; however, the performance may be limited in studies with small sample sizes (80). This may require more sophisticated constraints or assumptions to handle such issues. The small data size issue is also why we considered the built-in out-of-bag estimates (39) from the random forest model for cross-validation, which is shown to yield similar generalized accuracy to the leave-one-out cross-validation (81). This built-in feature prevents the need to further divide the data into smaller subsets for traditional k-fold cross-validation.

In summary, our results reveal fecal bacterial biomarkers can be utilized as indicators of consumption of specific foods

by healthy adults. These findings are important because they establish proof-of-concept for the use of bacterial biomarkers for predicting food intake. The constructed panel of microbes, or, more broadly, the analytic approach utilized to establish this panel, lays the foundational work necessary for future studies that aim to predict food intake using fecal microbial biomarkers. Furthermore, generation of biomarkers from stool samples has the potential to increase the robustness of other biomarkers from urine or blood to bolster objective assessment of dietary intake and compliance of participants in nutrition intervention studies. Overall, the identification of biomarkers of food intake will complement traditional dietary assessment methods, advance compliance evaluation in nutritional intervention studies, and move us toward diet–microbiota tailored therapies for disease prevention and treatment. Ultimately, establishing microbial biomarkers of food intake will facilitate the development of personalized nutrition approaches that modulate the gastrointestinal microbiome for human health benefit.

Acknowledgments

We thank Heather Guetterman, Jennifer Kaczmarek, Andrew Taylor, and Sharon Thompson for their technical assistance. The authors' responsibilities were as follows—DJB, JAN, CSC, NAK, and HDH: conceptualized the original study designs and oversaw the data collection and analyses; HDH, RZ, LSA, MEW, and CB: conceptualized the analysis approach for the present study; LMS, AM, and YL: analyzed the data; HDH, RZ, LMS, AM, YL, LSA, MEW, and CB: interpreted the data; HDH and RZ: have primary responsibility for the final content; and all authors: contributed to, read, and approved the final manuscript.

References

1. Marchesi JR, Adams DH, Fava F, Hermes GDA, Hirschfield GM, Hold G, Quraishi MN, Kinross J, Smidt H, Tuohy KM, et al. The gut microbiota and host health: a new clinical frontier. Gut 2016;65:330–9.
2. Knight R, Wegener Parfrey L, Ursell LK, Clemente JC. The impact of the gut microbiota on human health: an integrative review. Cell 2012;148:1258–70.
3. Holscher HD. Dietary fiber and prebiotics and the gastrointestinal microbiota. Gut Microbes 2017;8:172–84.
4. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science 2011;334:105–8.
5. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014;505:559–63.
6. 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;96:296–301.
7. Baer DJ, Gebauer SK, Novotny JA. Walnuts consumed by healthy adults provide less available energy than predicted by the Atwater factors. J Nutr 2016;146:9–13.
8. Zou ML, Moughan PJ, Awati A, Livesey G. Accuracy of the Atwater factors and related food energy conversion factors with low-fat, high-fiber diets when energy intake is reduced spontaneously. Am J Clin Nutr 2007;86:1649–56.
9. Holscher HD, Guetterman HM, Swanson KS, An R, Matthan NR, Lichtenstein AH, Novotny JA, Baer DJ. Walnut consumption alters the gastrointestinal microbiota, microbially derived secondary bile acids, and health markers in healthy adults: a randomized controlled trial. J Nutr 2018;148:861–7.
10. Holscher HD, Taylor AM, Swanson KS, Novotny JA, Baer DJ. Almond consumption and processing affects the composition of the gastrointestinal microbiota of healthy adult men and women: a randomized controlled trial. Nutrients 2018;10:126.
11. Thompson SV, Bailey MA, Taylor AM, Kaczmarek JL, Krug AR, Edwards CG, Reeser GE, Burd NA, Khan NA, Holscher HD. Avocado consumption alters intestinal bacteria abundance and metabolite concentrations among adults with overweight or obesity: a randomized, controlled trial. J Nutr 2020 Aug 17 (Epub ahead of print; DOI: 10.1093/jn/nxaa219).
12. Kaczmarek JL, Liu X, Charron CS, Novotny JA, Jeffery EH, Seifried HE, Ross SA, Miller MJ, Swanson KS, Holscher HD. Broccoli consumption affects the human gastrointestinal microbiota. J Nutr Biochem 2019;63:27–34.
13. Gozzi G, Cavallo N, Vannini L, De Angelis M, Di Cagno R, Gobbetti M, Maranzano V, Cosola C, Montemurno E, Gesualdo L. Effect of whole-grain barley on the human fecal microbiota and metabolome. Appl Environ Microbiol 2015;81:7945–56.
14. Subar AF, Freedman LS, Tooze JA, Kirkpatrick SI, Boushey C, Neuhouser ML, Thompson FE, Potischman N, Guenther PM, Tarasuk V, et al. Addressing current criticism regarding the value of self-report dietary data. J Nutr 2015;145:2639–45.
15. Shim J-S, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health 2014;36:e2014009.
16. Schatzkin A, Subar AF, Moore S, Park Y, Potischman N, Thompson FE, Leitzmann M, Hollenbeck A, Morrissey KG, Kipnis V. Observational epidemiologic studies of nutrition and cancer: the next generation (with better observation). Cancer Epidemiol Biomarkers Prev 2009;18:1026–32.
17. Lampe JW, Huang Y, Neuhouser ML, Tinker LF, Song X, Schoeller DA, Kim S, Raftery D, Di C, Zheng C, et al. Dietary biomarker evaluation in a controlled feeding study in women from the Women's Health Initiative cohort. Am J Clin Nutr 2017;105:466–75.
18. Brown IJ, Dyer AR, Chan Q, Cogswell ME, Ueshima H, Stamler J, Elliott P; INTERSALT Co-operative Research Group. Estimating 24-hour urinary sodium excretion from casual urinary sodium concentrations in Western populations: the INTERSALT study. Am J Epidemiol 2013;177:1180–92.
19. Scott TM, Rasmussen HM, Chen O, Johnson EJ. Avocado consumption increases macular pigment density in older adults: a randomized, controlled trial. Nutrients 2017;9:919.
20. 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;67:1205–14.
21. Bingham SA. Biomarkers in nutritional epidemiology. Public Health Nutr 2002;5:821–7.
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;10:1911.
23. Vázquez-Manjarrez N, Weinert CH, Ulaszewska MM, Mack CI, Micheau P, Pétéra M, Durand S, Pujos-Guillot E, Egert B, Mattivi F, et al. Discovery and validation of banana intake biomarkers using untargeted metabolomics in human intervention and cross-sectional studies. J Nutr 2019;149:1701–13.
24. Woodside JV, Draper J, Lloyd A, McKinley MC. Use of biomarkers to assess fruit and vegetable intake. Proc Nutr Soc 2017;76:308–15.
25. Münger LH, Garcia-Aloy M, Vázquez-Fresno R, Gille D, Rosana ARR, Passerini A, Soria-Florido MT, Pimentel G, Sajed T, Wishart DS, et al. Biomarker of food intake for assessing the consumption of dairy and egg products. Genes Nutr 2018;13:26.
26. Garcia-Aloy M, Hulshof PJM, Estruel-Amades S, Osté MCJ, Lankinen M, Geleijnse JM, De Goede J, Ulaszewska M, Mattivi F, Bakker SJL, et al. Biomarkers of food intake for nuts and vegetable oils: an extensive literature search. Genes Nutr 2019;14:7.
27. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, et al. American gut: an open platform for citizen science microbiome research. mSystems 2018;3:e00031–18.
28. US Food and Drug Administration. Specific Requirements for Nutrient Content Claims, 21 C.F.R. Chapter I, Subchapter B, Part 101, Subpart D(2020).
29. Edwards CG, Walk AM, Thompson SV, Reeser GE, Erdman JW, Burd NA, Holscher HD, Khan NA. Effects of 12-week avocado consumption

on cognitive function among adults with overweight and obesity. Int J Psychophysiol 2020;148:13–24.
30. 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;120:1370–9.
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: 406.1.
32. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 2016;13:581–3.
33. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 2019;37:852–7.
34. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. Nucl Acids Res 2014;42:D643.
35. Gebauer SK, Novotny JA, Bornhorst GM, Baer DJ. Food processing and structure impact the metabolizable energy of almonds. Food Funct 2016;7:4231–8.
36. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118–27.
37. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Statist Assoc 1952;47:583–621.
38. Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2:18–22.
39. Breiman L. Out-of-bag estimation. Technical report. Berkeley (CA): Statistics Department, University of California; 1996.
40. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2009.
41. Jackson JE. A user's guide to principal components. New York: John Wiley & Sons; 2005.
42. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 2001;45:171–86.
43. Qu K, Guo F, Liu X, Lin Y, Zou Q. Application of machine learning in microbiology. Front Microbiol 2019;10:827.
44. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, et al. Personalized nutrition by prediction of glycemic responses. Cell 2015;163:1079–94.
45. Fu J, Bonder MJ, Cenit MC, Tigchelaar EF, Maatman A, Dekens JAM, Brandsma E, Marczynska J, Imhann F, Weersma RK, et al. The gut microbiome contributes to a substantial proportion of the variation in blood lipids. Circ Res 2015;117:817–24.
46. Ertz-Archambault N, Keim P, Von Hoff D. Microbiome and pancreatic cancer: a comprehensive topic review of literature. World J Gastroenterol 2017;23:1899–908.
47. Das P, Marcišauskas S, Ji B, Nielsen J. Metagenomic analysis of bile salt biotransformation in the human gut microbiome. BMC Genomics 2019;20:517.
48. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med 2009;1:6ra14.
49. Raiten DJ, Namasté S, Brabin B, Combs GJ, L'Abbe MR, Wasantwisut E, Darnton-Hill I. Executive summary—Biomarkers of Nutrition for Development: building a consensus. Am J Clin Nutr 2011;94: 633S–50S.
50. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev 2011;35:343–59.
51. Hughes RL, Kable ME, Marco M, Keim NL. The role of the gut microbiome in predicting response to diet and the development of precision nutrition models. Part II: results. Adv Nutr 2019;10: 979–98.
52. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3.
53. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B 1996;58:267–88.
54. Auret L, Aldrich C. Interpretation of nonlinear relationships between process variables by use of random forests. Miner Eng 2012;35:27–42.
55. Gangopadhyay N, Hossain MB, Rai DK, Brunton NP. A review of extraction and analysis of bioactives in oat and barley and scope for use of novel food processing technologies. Molecules 2015;20:10884–909.
56. Jefferson A, Adolphus K. The effects of intact cereal grain fibers, including wheat bran on the gut microbiota composition of healthy adults: a systematic review. Front Nutr 2019;6:33.
57. Pennington JAT, Fisher RA. Classification of fruits and vegetables. J Food Compos Anal 2009;22:S23–31.
58. Lund ED, Smoot JM. Dietary fiber content of some tropical fruits and vegetables. J Agric Food Chem 1982;30:1123–7.
59. Houben K, Jolie RP, Fraeye I, Van Loey AM, Hendrickx ME. Comparative study of the cell wall composition of broccoli, carrot, and tomato: structural characterization of the extractable pectins and hemicelluloses. Carbohydr Res 2011;346:1105–11.
60. Satija A, Hu FB. Cardiovascular benefits of dietary fiber. Curr Atheroscler Rep 2012;14:505–14.
61. Li BW, Andrews KW, Pehrsson PR. Individual sugars, soluble, and insoluble dietary fiber contents of 70 high consumption foods. J Food Compos Anal 2002;15:715–23.
62. Duester KC. Avocado fruit is a rich source of beta-sitosterol. J Acad Nutr Diet 2001;101:404–5.
63. Han J-H, Yang Y-X, Feng M-Y. Contents of phytosterols in vegetables and fruits commonly consumed in China. Biomed Environ Sci 2008;21:449–53.
64. Dreher ML, Davenport AJ. Hass avocado composition and potential health effects. Crit Rev Food Sci Nutr 2013;53:738–50.
65. Abdel-Aal ESM, Akhtar H, Zaheer K, Ali R. Dietary sources of lutein and zeaxanthin carotenoids and their role in eye health. Nutrients 2013;5:1169–85.
66. Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, Kim AD, Shmagel AK, Syed AN, Walter J, et al. Daily sampling reveals personalized diet-microbiome associations in humans. Cell Host Microbe 2019;25:789–802.e5.
67. Duncan SH, Flint HJ, Sheridan PO, Scott KP, Gratz SW. The influence of diet on the gut microbiota. Pharmacol Res 2012;69:52–60.
68. Gorissen L, Raes K, Weckx S, Dannenberger D, Leroy F, De Vuyst L, De Smet S. Production of conjugated linoleic acid and conjugated linolenic acid isomers by *Bifidobacterium* species. Appl Microbiol Biotechnol 2010;87:2257–66.
69. Devillard E, McIntosh FM, Duncan SH, Wallace RJ. Metabolism of linoleic acid by human gut bacteria: different routes for biosynthesis of conjugated linoleic acid. J Bacteriol 2007;189:2566–70.
70. Sayin SI, Wahlström A, Felin J, Jäntti S, Marschall H-U, Bamberg K, Angelin B, Hyötyläinen T, Orešič M, Bäckhed F. Gut microbiota regulates bile acid metabolism by reducing the levels of tauro-beta-muricholic acid, a naturally occurring FXR antagonist. Cell Metab 2013;17:225–35.
71. Devkota S, Wang Y, Musch MW, Leone V, Fehlner-Peach H, Nadimpalli A, Antonopoulos DA, Jabri B, Chang EB. Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in *Il10*$^{-/-}$ mice. Nature 2012;487:104–8.
72. Martinez I, Lattimer JM, Hubach KL, Case JA, Yang J, Weber CG, Louk JA, Rose DJ, Kyureghian G, Peterson DA, et al. Gut microbiome composition is linked to whole grain-induced immunological improvements. ISME J 2013;7:269–80.
73. Vanegas SM, Meydani M, Barnett JB, Goldin B, Kane A, Rasmussen H, Brown C, Vangay P, Knights D, Jonnalagadda S, et al. Substituting whole grains for refined grains in a 6-wk randomized trial has a modest effect on gut microbiota and immune and inflammatory markers of healthy adults. Am J Clin Nutr 2017;105:635–50.
74. Bamberger C, Rossmeier A, Lechner K, Wu L, Waldmann E, Fischer S, Stark RG, Altenhofer J, Henze K, Parhofer KG. A walnut-enriched diet affects gut microbiome in healthy Caucasian subjects: a randomized, controlled trial. Nutrients 2018;10:244.
75. Carmody RN, Bisanz JE, Bowen BP, Maurice CF, Lyalina S, Louie KB, Treen D, Chadaideh KS, Maini Rekdal V, Bess EN, et al. Cooking shapes the structure and function of the gut microbiome. Nat Microbiol 2019;4:2052–63.
76. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Saitta L , editor. ICML'96: Proceedings of the Thirteenth International

Conference on Machine Learning. San Francisco (CA): Morgan Kaufmann Publishers; 1996. pp. 148–56.
77. Chen T, He T, Benesty M, Khotilovich V, Tang Y. Xgboost: extreme gradient boosting. R Package version 04-2. 2015. pp. 1–4.
78. Zhang C, Liu Y. Multicategory angle-based large-margin classification. Biometrika 2014;101:625–40.
79. Hassoun MH. Fundamentals of artificial neural networks. Cambridge (MA): MIT Press; 1995.
80. Zhou Y-H, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. Front Genet 2019;10:579.
81. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.