PNAS RESEARCH ARTICLE | APPLIED BIOLOGICAL SCIENCES

Diversity of plant DNA in stool is linked to dietary quality, age, and household income

Brianna L. Petrone[a,b], Ammara Aqeel[a], Sharon Jiang[a], Heather K. Durand[a], Eric P. Dallow[a], Jessica R. McCann[a], Holly K. Dressman[c], Zhengzheng Hu[c], Christine B. Tenekjian[d], William S. Yancy Jr.[d,e], Pao-Hwa Lin[f], Julia J. Scialla[g,h], Patrick C. Seed[i], John F. Rawls[a,j], Sarah C. Armstrong[k], June Stevens[l], and Lawrence A. David[a,j,1]

Edited by Jeffrey Gordon, Washington University in St. Louis School of Medicine, St. Louis, MO; received April 10, 2023; accepted May 10, 2023

Eating a varied diet is a central tenet of good nutrition. Here, we develop a molecular tool to quantify human dietary plant diversity by applying DNA metabarcoding with the chloroplast *trnL*-P6 marker to 1,029 fecal samples from 324 participants across two interventional feeding studies and three observational cohorts. The number of plant taxa per sample (plant metabarcoding richness or pMR) correlated with recorded intakes in interventional diets and with indices calculated from a food frequency questionnaire in typical diets (ρ = 0.40 to 0.63). In adolescents unable to collect validated dietary survey data, *trnL* metabarcoding detected 111 plant taxa, with 86 consumed by more than one individual and four (wheat, chocolate, corn, and potato family) consumed by >70% of individuals. Adolescent pMR was associated with age and household income, replicating prior epidemiologic findings. Overall, *trnL* metabarcoding promises an objective and accurate measure of the number and types of plants consumed that is applicable to diverse human populations.

food biodiversity | diet quality | nutrition | dietary assessment | high-throughput sequencing

Significance

The past 30 y have seen repeated calls for innovation in dietary assessment of human populations, yet field standard methods in epidemiology all still rely on asking individuals to self-report their diet. We developed a scalable tool for assessment of dietary plant intake in free-living humans by sequencing plant DNA in stool. Of many candidate summary metrics, we validated it for dietary plant diversity and used it to identify patterns across hundreds of individuals who varied by age, race, ethnicity, and income. In doing so, our work opens the door for use of breakthroughs in DNA sequencing to monitor and improve nutrition.

Unraveling the association between dietary patterns and health outcomes requires robust dietary assessment tools. Increasingly, researchers have turned to metabolic dietary biomarkers measured from biological specimens like blood, urine, and stool as alternatives to or validation for self-reported dietary data, which have well-characterized random and systematic errors (1–4). Dietary biomarkers have been developed and validated for total energy intakes (1), individual nutrients (1), food groups (5), and dietary patterns (6, 7). However, candidate biomarkers have failed to match both the resolution and breadth of information derived from self-reports: for example, the Diet History Questionnaire, a food frequency questionnaire (FFQ) developed for the general US population, collects data on frequency and intake amount of 135 individual food and beverage items (8). Because metabolites are derived from specific nutrients (limiting their generalizability) or transformed by the body (limiting their specificity) (9), no single metabolic biomarker can uniquely identify a comparable range of foods.

Genomic biomarkers, in contrast, are promising candidates for characterizing the full complement of foods that make up a dietary pattern. Genomic regions called "molecular barcodes," which identify a food species by its DNA sequence, can be amplified and sequenced from the residual pool of food-derived DNA in stool. Despite demonstrated utility in dietary studies in non-human species (10), such "DNA metabarcoding" approaches have only been applied to human samples in a restricted set of conditions: two fecal samples in the development of a very short plant barcode for amplification of highly degraded DNA (11); 54 fecal samples largely originating from restricted interventional diets (12); and 48 individuals' stomach contents collected during autopsy (13). Together, these studies demonstrated that DNA metabarcoding is possible from human digesta and stool, capable of identifying a range of known foods (47 plant taxa in ref. 12 and 33 plant and 9 animal taxa in ref. 13), may indicate compliance with dietary intervention (12), and has a high detection sensitivity for food items from a plant-based diet (86%) (12). However, only 20 of the sequenced stool samples described above originated from individuals consuming their typical, self-selected diets, the exposure of interest in observational nutritional studies. As a result, it is unknown whether DNA metabarcoding data capture variation in habitual dietary intake that can be leveraged for epidemiologic research.

Furthermore, logistical challenges remain to be overcome before DNA metabarcoding can be used as a dietary biomarker. Amplification of plant material from human stool has a high technical failure rate (50% of fecal samples in ref. 12), and bioinformatic analysis is not standardized. In addition to these methodological details, DNA metabarcoding

Preprint: https://doi.org/10.1101/2022.06.13.22276343.

Author contributions: B.L.P., C.B.T., W.S.Y., J.J.S., P.C.S., J.F.R., S.C.A., J.S., and L.A.D. designed research; B.L.P., S.J., H.K. Durand, E.P.D., J.R.M., H.K. Dressman, Z.H., and P.-H.L. performed research; B.L.P. and A.A. analyzed data; S.J., H.K. Durand, E.P.D., J.R.M., H.K. Dressman, Z.H., C.B.T., W.S.Y., P.-H.L., J.J.S., P.C.S., J.F.R., and S.C.A. supported collection, processing, or sharing of human samples; J.S. provided guidance and feedback; and B.L.P., A.A., S.J., H.K. Durand, J.R.M, P.-H.L., J.J.S., J.F.R., J.S., and L.A.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: lawrence.david@duke.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2304441120/-/DCSupplemental.

Published June 27, 2023.

data have not been collected at a scale necessary to power human nutrition studies, which typically include hundreds if not thousands of participants (4).

Beyond the challenges of generating DNA metabarcoding data at scale, there are also conceptual hurdles for analyzing it. The number of potential species identified by DNA metabarcoding is comparable in complexity to other marker gene methods like 16S rRNA microbiome sequencing and to established self-report dietary assessments, which describe food intakes according to hundreds of taxa or nutrient features, respectively (14). Strategies for analyzing these datasets have evolved over decades and are still a focus of active development and revision (15, 16). As a result, how the hundreds of dietary species that could be measured by DNA metabarcoding should be analyzed to characterize a population or study relationships to health outcomes in humans remains unaddressed. In animals, dietary species abundances have been used to calculate within- and between-sample diversity (10), identify relationships to nutritional status (17), and quantitatively estimate the biomass of individual source foods in the diet (18). None of these approaches has yet been tried in human data, and results from animals may not translate to human diets, which include foods that are cooked, prepared, processed, or stored prior to eating.

To address the need for broadly scoping biomarkers of food intake with practical utility in nutrition research, we develop here DNA metabarcoding with the plastid *trnL*-P6 marker (11) as an epidemiological tool for assessing dietary plant intake in humans. We first report several protocol adaptations that make *trnL* metabarcoding more reliable in human stool samples. Next, to understand whether 1) DNA-based dietary data are accurate and useful in an epidemiologic context and 2) how exactly these data should be utilized, we apply plant DNA metabarcoding to 1,029 fecal samples from 324 participants, a scale comparable to validation studies in nutritional epidemiology. We assess the per-plant, quantitative accuracy of *trnL* metabarcoding in a cohort with high-quality accompanying dietary data, as well as the feasibility and validity of dietary diversity as a summary measure of intake. Finally, we show that *trnL* metabarcoding-derived dietary diversity metrics can be valuable for testing epidemiological hypotheses and for exploratory analysis of cohorts with limited dietary data.

Results

***trnL* Metabarcoding Protocol Development.** We developed a molecular approach for measuring dietary plant intake by amplifying and sequencing residual DNA from stool samples using the *trnL*-P6 region of the chloroplast genome ("*trnL* metabarcoding"; Fig. 1*A*). To scale *trnL* metabarcoding for population-level applications, we followed recommendations for microbiome metabarcoding studies (19) to refine our prior protocol (12), which was capable of detecting dietary plant taxa but limited by a low PCR success rate (~50%). We moved from a standard- to a high-fidelity polymerase to reduce PCR errors and facilitate accurate taxonomic assignment of *trnL* sequence variants and switched from a one-step to a two-step amplification protocol to avoid bias from barcode differences in the primary amplification and reduce primer synthesis costs. In the primary amplification, we adjusted reaction annealing temperature from 55 to 63 °C to reduce formation of nonspecific products and maximize amplification and sequencing yields (*SI Appendix*, Fig. S1 *A–C*), which further improved when template volume was quadrupled (*SI Appendix*, Fig. S1 *D–F*). In the second amplification step, an increase from 8 to 10 barcoding PCR cycles improved yields a further 3.4-fold. In samples tested pre- and postoptimization, these changes collectively resulted in median increases of 2.2 ng/μL in amplified DNA, 18,600 *trnL* sequencing reads, and four additional plant taxa detected per sample (n = 199, *SI Appendix*, Fig. S1 *G–I*). Overall, the optimized protocol had a PCR success rate of 92%, which exceeds response rates (76 to 82%) for validated FFQs (20), which are widely used instruments for the collection of self-reported dietary intake.

Bioinformatically, our optimization focused on assigning a plant taxon to *trnL* sequencing reads using an expanded reference database. In our previous work, 27% of reads did not have an exact match to a reference sequence and could not be included in subsequent analyses. We therefore expanded our reference database of dietary plants from 185 sequences (representing 72 species) to 791 sequences (468 species). In comparison to a recent food plant phylogeny (21), these species included members of 62% of all families and 83% of major crop families tracked by the Food and Agriculture Organization (FAO) (Fig. 1*B*). A curated, food-specific reference database also significantly improved the reported resolution of the *trnL*-P6 loop (11): of all possible sequences identified by the reference, a plant species, genus, or family could be exactly specified in 83.2%, 92.6%, or 99.3% of cases, respectively. To reduce the percentage of reads without a taxonomic assignment, we shifted from grouping similar DNA sequences into operational taxonomic units to inferring exact amplicon sequence variants (ASVs), which enabled calculation of common summary metrics (*i.e.*, within-sample or alpha diversity) independent of a reference assignment and straightforward merging of metabarcoding datasets from multiple sequencing runs that use the same primers and analysis parameters (22). In concert, our expanded reference database and revised pipeline reduced the percentage of unassigned reads to 0.9% (per-sample median 0.2% and median absolute deviation 0.3%; *SI Appendix*, Fig. S2).

***trnL* Metabarcoding Validation Using Interventional Diets.** With our optimized experimental and bioinformatic protocols, we examined the potential for *trnL* metabarcoding to capture recorded intake of specific plant taxa in a cohort of individuals undergoing a dietary intervention ("Weight Loss"; Table 1; n = 41 samples from four individuals). Members of the Weight Loss cohort were clients of a residential-style, medically supervised weight loss center, with all their weekday meals prepared in the center's cafeteria and consumed on site. Although Weight Loss diets were interventional, they had high day-to-day variability and included a large number of unique items. Meal recipes and participant orders were logged by a digital menu system, which allowed us to specify a plant taxon for 96% of the 425 unique plant-derived food items consumed, including those from complex meals (*e.g.*, "Mushroom Wild Rice Pilaf" could be separated into wild rice, white rice, portobello mushrooms, onion, pecans, thyme, parsley, and sage).

We began by assessing the detection accuracy of *trnL* metabarcoding for individual plant taxa. Overall, 76% of taxa in participants' cafeteria records were also present in *trnL* metabarcoding data (n = 56 of 74; *SI Appendix*, Fig. S3*A*). Comparing each Weight Loss stool sample to corresponding menu records from 2 d prior to collection, *trnL* and menu data were consistent in 74% of per-sample, per-taxon comparisons (14.4% true positives and 59.7% true negatives; *SI Appendix*, Fig. S3*B*). Still, concordance between *trnL* and menu data varied across plants (n = 96; *SI Appendix*, Fig. S3*B*): per-taxon accuracy, or the percentage of detections that concorded with the menu, ranged from 97 to 9% (median 78%, absolute deviation 21%).

One factor that can markedly impact perceived accuracy is the reliability of the "known" diet record. There is no gold standard method of dietary assessment, and Weight Loss center menus were an imperfect comparator despite their detail: menus indicated only

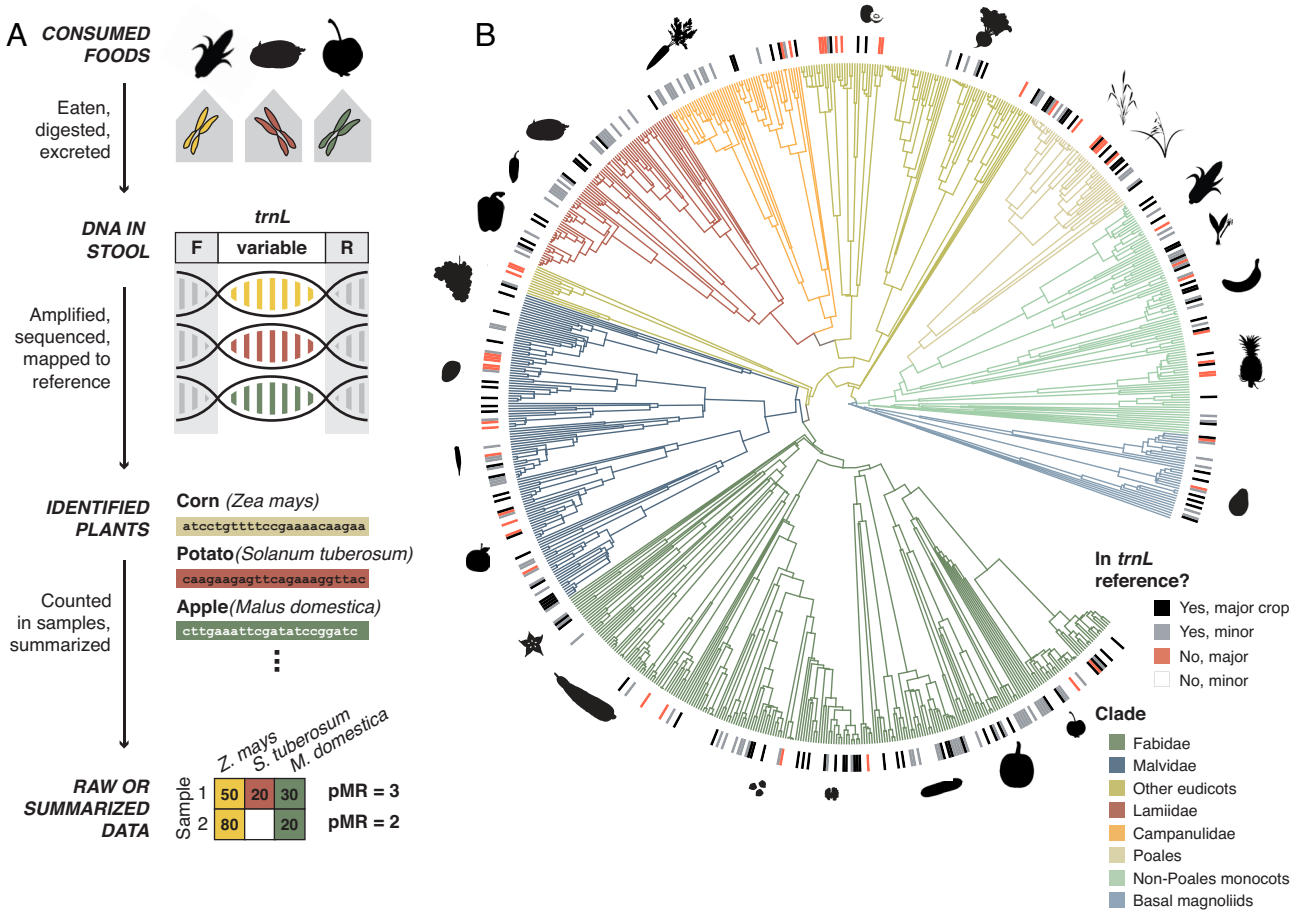

Fig. 1. Generation and scope of *trnL* metabarcoding data for dietary plant intake. (*A*) Conceptual overview of *trnL* metabarcoding protocol and pMR calculation. Conserved primers (F and R) flank a variable *trnL* region, allowing amplification of a mixed pool of plant food-derived DNA from stool. Following sequencing and taxonomic identification, data can be analyzed as the presence or count of each plant taxon per sample or metrics like the number of taxa per sample (pMR). (*B*) The reference *trnL* sequence database used for taxonomic assignment had broad representation (black and gray tick marks in the outer ring) of food crop species [full phylogenetic tree (21)] and included multiple sequences for 27% of plant taxa, which indicates within-food genetic variation at the *trnL*-P6 locus. Leaves in the crop tree terminate at the species level, although 70 subspecies- and 52 variety-level taxa were included in the full reference. Plant crops tracked by the FAO ("major") were more likely to be included in the reference than untracked crops ("minor"; Chi-square 188.94, df = 2, $P < 10^{-15}$). Example plants from each clade are shown in silhouette. Clockwise from legend, these are apple, pumpkin, cucumber, walnut, chickpea, cassava, starfruit, orange, okra, mango, grape, bell pepper, chili pepper, potato, carrot, kiwi, beet, rice, wheat, corn, onion, banana, pineapple, and avocado.

ordered foods rather than consumed ones and participants could eat foods that were not tracked by the menu system (*e.g.*, a daily fruit offering, salad bar, or off-menu eating that occurred away from the center). To specifically evaluate how features of the Weight Loss diet record may have affected per-taxon detection accuracy, we examined a stricter intervention cohort consuming a Western-style diet that was provisioned food from four recurring daily menus, with uneaten items returned and detailed daily food logs captured additional items consumed ("Controlled Feeding," Table 1, $n = 28$ samples from 14 individuals) (23). Across foods in common between the Weight Loss and Controlled Feeding cohorts ($n = 42$), detection errors were significantly correlated for false negative rate (Spearman $\rho = 0.63$, $P = 0.0003$; *SI Appendix*, Fig. S3*C*). Consistency in false negatives (*trnL* metabarcoding failing to detect a food recorded in the menu) supports a model in which food-specific factors affecting detection like digestibility and chloroplast copy number remain the same across study settings. In contrast, false-positive rates were uncorrelated between studies (Spearman $\rho = -0.14$ and $P = 0.42$; *SI Appendix*, Fig. S3*C*). This lack of relationship for false positives (*trnL* reporting a food taxon that the menu does not) is consistent with a factor that impacts false-positive detections in only one study: here, the fact that the Weight Loss menus have potential omissions.

We next examined the potential for DNA metabarcoding to capture variation in portion size in the Weight Loss cohort (by design, the Controlled Feeding menu did not feature large variation in servings of plant-based foods). Considering all foods in the Weight Loss cohort, the proportion of total *trnL* sequencing reads per sample was significantly associated with both continuous and categorical measures of portion size recorded in menu data (gram weight or intake tertile of food consumed, with Spearman $\rho = 0.31$, $P < 10^{-15}$ and two-tailed Mann–Whitney *U* test $P < 10^{-15}$, respectively; *SI Appendix*, Fig. S4 *A* and *B*). This quantitative signal persisted for a subset of individual taxa when *trnL* sequencing reads were compared to menu data from the 1 to 2 d before sampling on a per-food basis ($P < 0.05$ for grains (rye and wheat) and berries (strawberries and blackberries) and $P < 0.1$ for oats, blueberry, brassicas, celery, eggplant, mango, peas, peppers, raspberry, soy, tomato, and pommes (apples and pears), one-tailed Mann–Whitney *U* test with Benjamini–Hochberg correction; *SI Appendix*, Fig. S4*C*). These results suggest that *trnL* can provide quantitative information on portion size for select taxa.

***trnL* Metabarcoding Within-Sample Diversity as a Summary Metric.** Since our data indicated quantitative, but also heterogeneous, concordance between DNA metabarcoding and menu data at the

Table 1. Baseline characteristics of *trnL* metabarcoding cohorts

	Weight Loss	Controlled Feeding	Adult-1	Adult-2	Adolescent
n					
Individuals	4	14	28	32	246
Samples/ individual	11.5 ± 2.2	2 ± 0.0	16.0 ± 3.0	6.0 ± 0.0	2.0 ± 0.0
Total samples	41	28	387	189	384
Diet					
Type	Interventional reduced calorie	Western-style controlled feeding	Typical diet with fiber supplement	Typical diet with fiber or placebo snack bar	Typical diet
Assessment	Digital menu system	Provisioned intake plus reported items	FFQ (NCI DHQ3)	FFQ (NCI DHQ3)	Custom survey
Demographics					
Age, years	58.5 ± 8.8	67.9 ± 7.8	33.3 ± 12.0	25.6 ± 5.2	13.3 ± 2.3
Sex, % female	50	64	39	59	60
Race, %					
Black	0	29	4	3	53
White	75	57	68	44	38
Asian	0	0	11	38	2
Amer. Indian/ Alaska Native	25	7	0	0	0
Multiple	0	7	7	9	7
Ethnicity, % Hispanic	0	7	11	13	18
Health					
BMI	35.5 ± 4.6	29.6 ± 3.9	24.7 ± 2.4	22.9 ± 2.3	31.8 ± 10.1

All values are reported as mean ± SD except samples per individual, which is given as median ± median absolute deviation. Entries for Adult-1 and Adult-2 race do not sum to 100% due to missing raw data (*i.e.*, individuals that did not indicate a response). BMI, body mass index.

level of individual plants, we investigated whether an ensemble picture of intake might yield a more accurate whole-diet overview. Thus, we shifted to calculating within-sample diversity as a summary measure of diet.

We selected richness, or the number of plant taxa per sample, as a diversity metric. Dietary richness is both readily interpretable and consistent with the recommendation to "eat a variety of foods," which is a well-known component of public health nutrition guidance (24). In the early 1990s, studies conducted in the United States and globally demonstrated that counts of individual foods or food groups could serve as proxies for nutrient adequacy (25–27) or overall health (28); these findings prompted the introduction of dietary diversity–specific measurement tools by the FAO (29). Since the 2000s, diversity metrics have continued to be applied in studies of the relationship between overall dietary patterns and health outcomes (30–32) and have generated renewed interest as the interplay between food production, local ecosystem biodiversity, and the environment has become an area of policy focus (33).

We calculated plant metabarcoding richness (pMR, the number of plant taxa detected per sample with *trnL* metabarcoding) of samples in the Weight Loss cohort. On a per-sample basis, pMR was positively correlated with the average number of plants in menus from the 2 d prior to stool collection (Spearman $\rho = 0.42$, $P = 0.03$, Fig. 2*A*), which was selected as a comparison interval based on variation in interindividual gastrointestinal transit time around a median of 28 h (34, 35). pMR was unrelated to the menu data from the 2 d following sample collection or two randomly selected consecutive menu days (*SI Appendix*, Fig. S5). Outliers in Fig. 2*A* (defined in *Materials and Methods*) originated from the only participant that did not complete study surveys about off-menu eating, which may explain why pMR appeared to overestimate recorded intake in this participant. We also piloted two other within-sample summary metrics, Shannon diversity (which weights the diversity estimate by *trnL* read count) and Faith's phylogenetic diversity (which weights by the evolutionary relatedness of detected foods). The observed associations were consistent with those obtained for pMR ($\rho = 0.45$ and $P = 0.02$, Shannon; $\rho = 0.38$ and $P = 0.04$, Faith's). Still, we decided to focus on pMR in subsequent analyses because species richness is both simpler to understand and more akin to prior epidemiological measures of dietary diversity (26, 32) than Shannon or Faith's phylogenetic diversity. Overall, in a small, controlled setting with high-quality dietary data available for comparison, our data indicated that *trnL* diversity measured from stool was related to recorded dietary plant diversity.

***trnL* Metabarcoding Validation for Typical Diets.** Because the Weight Loss cohort consisted of a limited number of participants consuming an interventional health-promoting diet and dietary diversity measurement is of importance across the real-world spectrum of intakes, we next evaluated pMR as a biomarker of dietary diversity in a larger number of individuals eating their typical diet. We performed *trnL* metabarcoding on fecal DNA from two larger adult cohorts that were recruited for studies testing the impact of fiber supplementation ("Adult-1" and "Adult-2," $n = 28$ and $n = 32$, respectively; Table 1) (37, 38). Participants ate their typical diets, which were surveyed with an FFQ, a standard dietary assessment tool in nutritional epidemiology. Multiple stool samples were collected per week over a 6- or 3-wk period

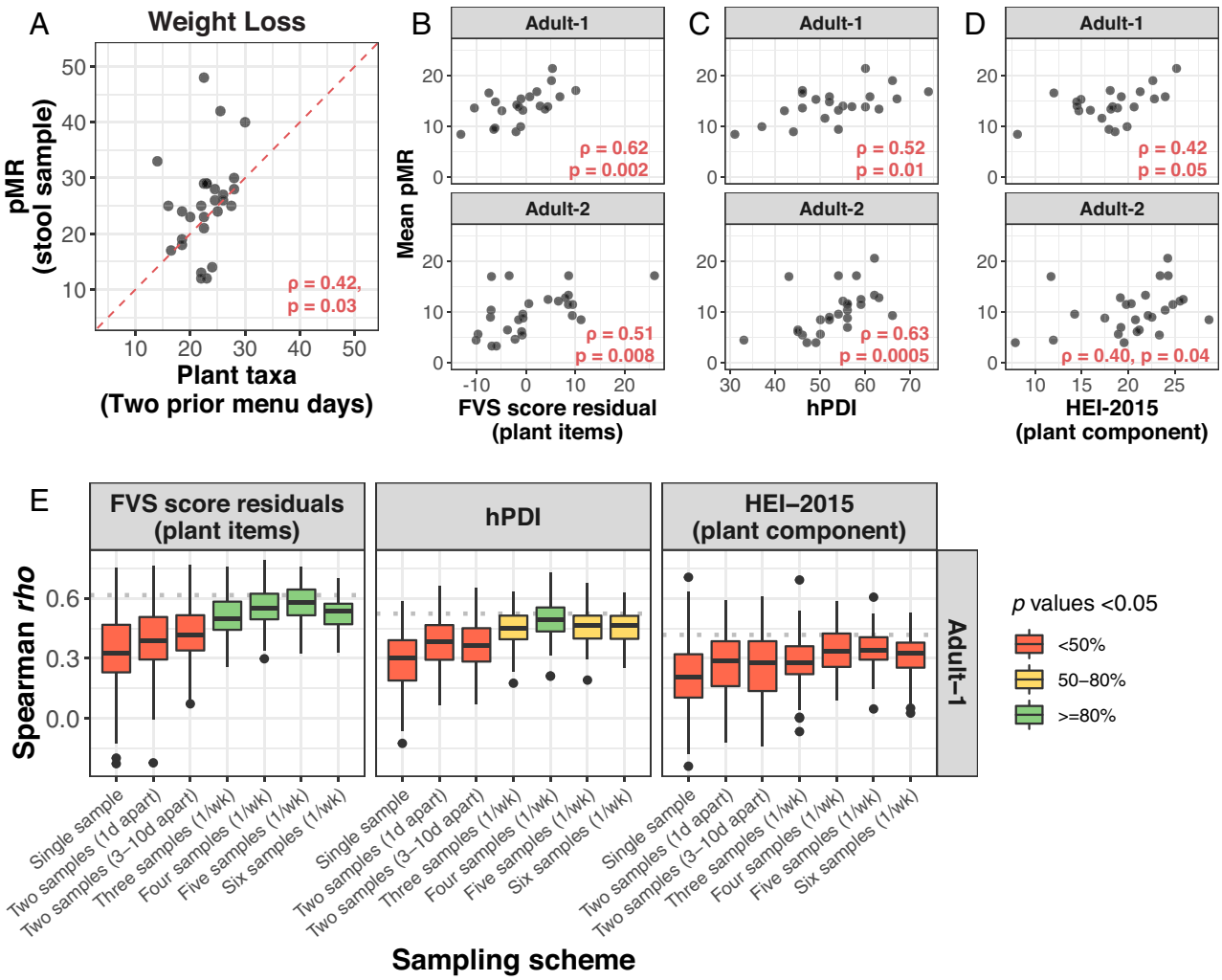

Fig. 2. pMR is associated with independent measures of dietary diversity and quality. (*A*) Correlation between pMR and number of plant taxa from recorded menus of Weight Loss participants from the 2 d prior to stool collection. The red dotted line denotes a theoretical perfect correspondence between the two measures. (*B–D*) Correlations between mean pMR (pMR averaged across all available stool samples per participant) and dietary diversity (*B*) and quality (*C* and *D*) indices derived from FFQ data in Adult-1 and Adult-2 participants. (*E*) Correlations from upper panels of (*B–D*) retested under candidate sampling schemes with mean pMR derived from a smaller number of stool samples. The "two samples (3 to 10 d apart)" is the current dietary assessment protocol used by the National Health and Nutrition Examination Survey (36). All boxplots represent ~100 random subsamples at each strategy, and color indicates the percentage of iterations reaching the statistical significance threshold of $P < 0.05$. Spearman correlations are two-tailed. FVS, Food Variety Score; hPDI, healthy plant-based dietary index; HEI-2015, Healthy Eating Index 2015.

(median of 16 and six samples/person for Adult-1 and Adult-2, respectively).

To assess the degree of alignment between pMR and FFQ data, we summarized reported foods into commonly used dietary indices. We included a dietary diversity score, which counted unique food items, and two dietary quality scores, which weighted food items or groups based on amount consumed and health benefit or harm. For a diversity index, we selected the Food Variety Score (FVS), a score with a 20-y history (26) that correlates with nutrient adequacy (25) and reduced risk of coronary heart disease and all-cause mortality (39). For quality scores, we evaluated the Healthy Eating Index 2015 (HEI), which indicates adherence to US dietary guidelines, and the healthy, unhealthy, and overall plant-based dietary indices (hPDI/uPDI/PDI), which assess plant presence and quality in the diet. We chose HEI and hPDI/uPDI because they have been previously linked to reduced risk of chronic disease morbidity or mortality (40–43) and demonstrated to correlate tightly ($\rho > 0.7$) with predictions based on microbiome composition (44), which, like pMR, is a stool-based molecular measurement. Because HEI and FVS both include animal components, we used only the portion of the score that referenced plants, and because FVS scores scaled linearly with dietary calories, we used energy-adjusted residuals in place of the raw score (*Materials and Methods*).

We identified significant, positive correlations between pMR and FFQ-based dietary indices in both Adult-1 and Adult-2 cohorts. Mean pMR per participant was positively correlated specifically with the plant component residuals of the FVS (Fig. 2*B*; Adult-1 Spearman $\rho = 0.62$, $P = 0.002$, Adult-2 $\rho = 0.51$, $P = 0.008$); with the healthful component of the PDI (hPDI) (Fig. 2*C*; Adult-1 $\rho = 0.52$, $P = 0.01$, Adult-2 $\rho = 0.63$, $P = 0.0005$); and with the plant-based component score of the Healthy Eating Index 2015 (HEI-2015) (Fig. 2*D*; Adult-1 $\rho = 0.42$, $P = 0.05$, Adult-2 $\rho = 0.40$, $P = 0.04$). Correlations were absent or negative when tested against animal-based or unhealthy component scores alone (*SI Appendix*, Fig. S6). All results except that for HEI-2015 were robust to rarefaction, a statistical downsampling to estimate richness in samples of varying sequencing depth (see additional details in *Materials and Methods*, *SI Appendix*, Table S2). These findings indicate that pMR, a molecular dietary diversity measure, can rank

individuals in a significantly similar way to multiple validated diversity and quality indices based on FFQ data, even though distinct sources of error underly the two forms of dietary assessment.

Given the dense stool sampling protocols of the Adult-1 and Adult-2 cohorts, we next sought to determine the minimum number of samples per participant necessary to capture a comprehensive view of dietary plant diversity. We generated collector's curves, an ecological tool used to assess richness as a function of sampling effort, for each participant (*SI Appendix*, Fig. S7). Unlike the average pMR calculated above, collector's curves provide a running tally of the number of unique plant taxa detected as samples from the same individual are successively pooled. Curves were well fit by a logarithm function, indicating that cumulative pMR plateaus with sufficient sampling. Consistent with prior work in diet records, which detected a plateau in "food repertoire" after 10 to 15 d of recorded intake (45), the early plateau phase was often reached by individuals with >15 stool samples in the Adult-1 cohort, but rarely in Adult-2 participants, who collected at most six samples.

Even though a dozen stool samples may be required to observe an individual's total potential dietary diversity, we found that averaged pMR from fewer samples could still reproduce the significant associations with dietary indices described above. Subsampling each participant recapitulated the significant correlation with hPDI at least 80% of the time and with FVS at least 50% of the time under at least one reduced sampling strategy in both cohorts (100 iterations at each strategy, unless fewer unique combinations were possible; Adult-1 in Fig. 2*E*; Adult-2 in *SI Appendix*, Fig. S8 due to more limited subsampling). The relationship to the HEI-2015 plant component score was not robust to subsampling, likely because it measures adequate intake of only five highly summarized food categories (*e.g.*, "total vegetables") and thus is better approximated by average pMR derived from larger number of samples. These results indicate that pMR from as few as three samples per person approximates the ranking of individuals by both a traditional dietary diversity index (FVS) and a dietary quality index (hPDI).

***trnL* Metabarcoding in Settings without Available Dietary Data.** We next applied DNA metabarcoding in a setting where traditional dietary assessment measures were not collected. In a pediatric study of gut microbiota in adolescents with and without obesity from racially, ethnically, and socioeconomically diverse backgrounds ("Adolescent," n = 246, 79% with BMI >95th percentile, 53% Black, 18% Hispanic, and >40% with household income <$50,000/year; Table 1), dietary assessment was limited to a custom 7-question survey. Two lengthier assessments had been eliminated within the first 10 enrolled participants as they proved too cumbersome for families to complete.

Because no data on specific food items were available for the Adolescent cohort, we leveraged the taxon-level food identifications of *trnL* metabarcoding to identify plants included in participant diets. Across the cohort, we detected 111 unique *trnL* sequence variants, which came from 46 plant families, 85 plant genera, and 72 plant species (the number of genera identified is larger than the number of species because certain plant taxa can only be identified with genus-level precision; see *Materials and Methods*). The most frequently observed food items were wheat or rye (the detected *trnL* sequence variant being the same for both foods; 96% of participants), chocolate (88%), corn (87%), and members of the potato family (a *trnL* sequence variant shared by potato, tomatillo, tamarillo, goji berry, cutleaf groundcherry, and edible nightshades; 71% of participants). However, the vast majority of *trnL* sequence variants had low prevalence across the pool of subjects, indicating a small set of commonly consumed foods and many more that were unique to the diets of only a handful of individuals (*SI Appendix*, Fig. S9*A*). The types of plant foods detected in the Adolescent cohort did not differ widely from the Weight Loss, Adult-1, or Adult-2 cohorts (*SI Appendix*, Fig. S9*B*). The calculated pMR ranged widely across the cohort (median 12 plant taxa per sample, median absolute deviation 4.4; Fig. 3*A*), and the presence or absence of foods in the diet indicated a spectrum of intakes rather than a partitioning of distinct dietary patterns (*SI Appendix*, Fig. S9*C*).

To test the utility of pMR as an epidemiologic tool, we explored whether we could replicate previously reported associations between dietary diversity and demographic variables. This was uniquely possible in the Adolescent cohort due to its large size, breadth of demographic data, and higher racial and ethnic diversity in comparison to the Weight Loss, Adult-1, and Adult-2 cohorts. In a multiple regression, Adolescent pMR was negatively associated with age ($\widehat{\beta}$ = -3.9 [95% CI −7.1 to −0.7], P = 0.02), positively associated with higher income categories ($\widehat{\beta}$ = 2.2 [0.03 to 4.4], P = 0.05 for $25,000 to 49,000/year and $\widehat{\beta}$ = 3.0 [0.1 to 5.9], P = 0.04 for $100,000/year, both relative to lowest bracket of <$25,000/year), trended higher with obesity status ($\widehat{\beta}$ = 1.8 [−0.1 to 3.7], P = 0.06), and lower with food insecurity ($\widehat{\beta}$= −1.7 [−3.7 to 0.3], P = 0.09), and was unrelated to sex, race, and ethnicity (all with P > 0.28; fitted coefficients in Fig. 3*B*, raw relationships in Fig. 3*C*). The negative association between pMR and age is consistent with data that American adolescents are less likely to eat dinner with their families as they age, which is associated with lower intakes of fruit, vegetable, and whole-grain foods (46). The positive trend between pMR and obesity status supports a recent recommendation to reduce emphasis on dietary diversity in adult populations (31) given inconsistent associations with lower adiposity (47, 48) and positive relationships to total energy intake, obesity, or body fat percentage detected in some studies (39, 49, 50). Finally, the positive association between pMR and higher income categories aligns with epidemiologic studies within and outside the United States, which report increased dietary diversity in households with higher socioeconomic status (25, 28, 51, 52). We performed a comparable literature review for covariates for which we did not detect a significant association and found both concordant [ethnicity (53)] and discordant [sex (28, 39) and race (39, 54)] results, although all were derived from nonadolescent cohorts. Thus, dietary plant diversity measured by pMR replicates the majority of known epidemiological findings from studies that used self-report-based diversity measures.

Discussion

In this study, we establish that retrospective dietary assessment with *trnL* metabarcoding applied to human stool is 1) readily applied at epidemiologic scale (100s to 1,000s of samples) with improved experimental and bioinformatic methods and 2) validated for measurement of the number of plant foods in the diet. This work, in which all but 18 of the 324 participants were consuming their typical diets, provides a large-scale application of *trnL* metabarcoding to individuals with self-selected diets, who are the population of interest for measurement of dietary exposures. In total, we detected 187 unique *trnL* sequence variants representing 146 taxa from 73% of major food crop families present in the expanded reference database; prior *trnL* metabarcoding studies conducted in humans detected at most 47 plant taxa (11–13).

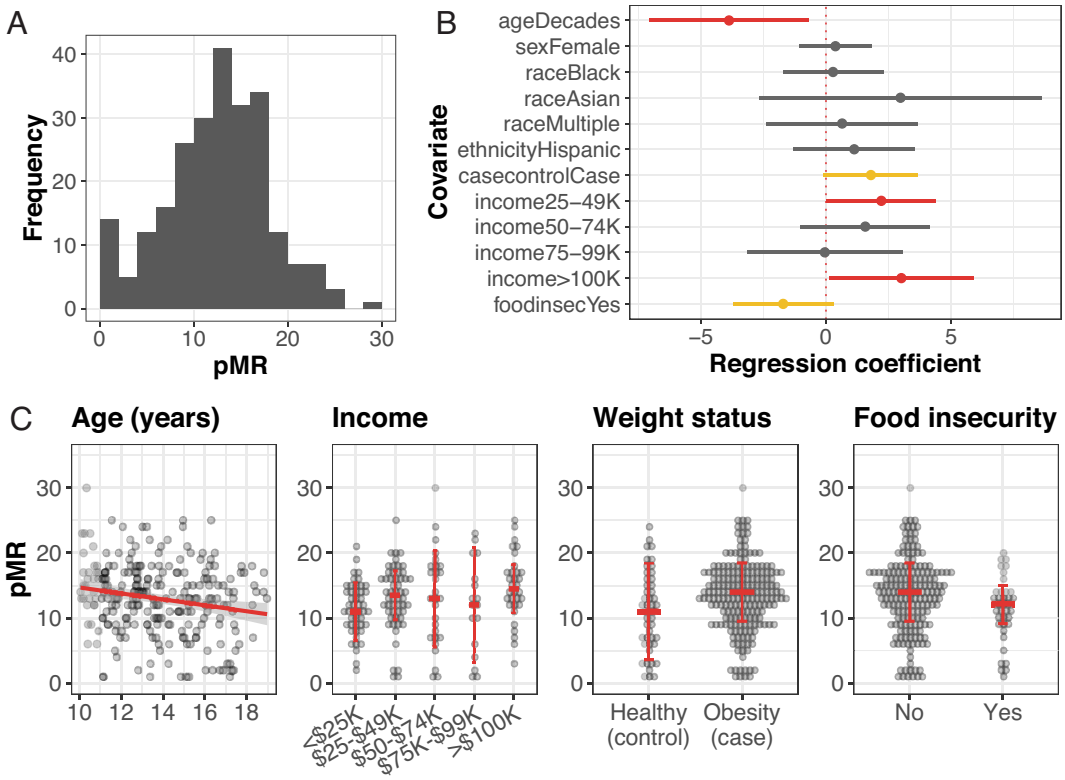

Fig. 3. pMR detects known relationships between dietary diversity and demographic, health, and socioeconomic variables. (*A*) Histogram of pMR across Adolescent samples. (*B*) Visualization of linear model output, showing effect sizes and 95% CIs of associations of demographic, clinical, and socioeconomic covariates with pMR as a response variable. Coefficient estimates with $P \leq 0.05$ are indicated in red, $0.05 < P \leq 0.1$ in yellow, and $P > 0.1$ in gray. For categorical variables shown, the reference category is as follows: white for race, non-Hispanic ethnicity for ethnicity, control (healthy body weight) for case–control status, self-reported income <$25,000 annually for income, and no occurrence of food running out for food insecurity. (*C*) Raw data underlying significant or trending covariates from (*B*).

Together, our results lay out the present uses and future potential of metabarcoding as a research tool in nutritional epidemiology. First, the strength of positive correlations (Spearman ρ ~ 0.4 to 0.6) between pMR and indices calculated from a validated self-report tool falls within a range used by epidemiologists to validate simple dietary assessments against more comprehensive ones. For example, the use of a simplified survey specifically for dietary diversity measurement is supported by Pearson correlations to nutrient adequacy from a more complex tool ranging from $r = 0.3$ to 0.6 (25, 27, 55). This means pMR can complement self-report tools by providing an alternative means of assessment in populations for whom validated dietary assessments do not exist or in settings where self-reports are limited by participant burden or resource constraints, as in the Adolescent cohort. Second, given the broadly similar magnitude of correlation with both dietary diversity (FVS) and quality (hPDI, HEI) indices (Fig. 2*B*), pMR may combine elements of the two by implicitly reflecting food features like amount and quality (*i.e.*, due to a detection bias in favor of items consumed in large quantities or without DNA degradation from industrial processing). Such a metric may allow researchers to develop a more robust definition of healthy dietary diversity. Teasing apart healthy versus unhealthy dietary diversity would be useful given recent concern that strict definitions of dietary diversity do not discriminate against unhealthy eating: in some studies of adults not at risk of nutrient inadequacy, increased dietary diversity associates with greater intake of processed foods, refined grains, and sweetened drinks (56). Third, unlike dietary self-reports, *trnL* metabarcoding provides data in the conserved language of DNA sequence, which overcomes challenges of manual food item identification, grouping, and nomenclature (57) and permits immediate harmonization of data across global studies.

No dietary assessment technique is universal in scope and without limitations. pMR does not reflect processing or cooking techniques used to prepare foods. Some FFQs and existing dietary diversity measures similarly do not consider food preparation (33) or do not count prepared items toward the final score [e.g., "Oils and fats" and "Sweets" are omitted from the Women's Dietary Diversity Score (29)]; as is, pMR correlates with measures of healthy eating that incorporate food preparation (Fig. 2*B*). Our definition of pMR as the number of plant food taxa per sample differs from conventional dietary diversity, which is calculated as number of foods or food groups over a reference period (usually 24 h). Due to interindividual differences in gastrointestinal transit time, pMR could summarize food intake over a period from 24 h to multiple days and aligned best with Weight Loss menu data when considering two prior days of intake. Pairing *trnL* metabarcoding with transit time indicators (edible dyes or proxies like Bristol stool scale scores) to perform an individual-specific adjustment may reveal even more robust associations between pMR and self-reported diversity metrics. A sampling period >1 d could even be advantageous for epidemiologic applications: multiple administrations of a 24-h dietary recall (*i.e.*, >1 d of diet) are recommended for examining associations to health outcomes (3). With 99.1% of reads in the data presented here assigned to a food taxon, we estimate that the *trnL* reference database provides nearly complete coverage of foods consumed in Western diets, but its performance for global cohorts is likely lower and will be the target of future updates. Finally, we have characterized only plant diversity, omitting animal and fungal sources of dietary variation. Existing

DNA barcodes for these kingdoms [12SV5 for animals (13) and ITS for fungi (58)] can be adapted as we have done here but require additional optimization for human dietary studies. Animal markers also amplify human DNA, which can be reduced by the addition of a human-specific blocking primer, and fungal markers may be prone to signal competition between dietary fungi and the resident gut mycobiome. Candidate universal markers like the nuclear 18S rRNA gene (59) might even be capable of simultaneously detecting important food taxa across all three kingdoms.

These findings position *trnL* metabarcoding as a candidate genetic biomarker of plant intake and support its use to derive pMR prospectively from any individual able to provide a stool sample and retrospectively from biorepositories or DNA extracted for another purpose (*e.g.*, 16S rRNA or metagenomic sequencing). Our *trnL* metabarcoding experimental protocols, bioinformatic pipeline, and reference database are publicly available and actively maintained. Estimated costs per sample are comparable with 16S sequencing and competitive with diet survey costs for an interviewer-administered 24-h recall. *trnL* metabarcoding therefore has the potential to be as widely available and accessible as 16S sequencing technology is for gut microbiome profiling from stool samples. With this tool in hand, we can envision studies that use *trnL* metabarcoding to develop a unified framework for the impact of dietary diversity on health, which has been hampered by the enormous range of edible plant species [>4,000 compiled from published records (60)] and lack of standardization in underlying assessment methodology (26). We can also contemplate using *trnL* metabarcoding to tally food sources in global contexts without needing to prioritize by commercial or cultural importance, prevalence as common or rare, or status as domesticated or wild. In this form, dietary data could be directly connected to environmental biodiversity monitoring (33) or food system sustainability (61). Thus, DNA metabarcoding can become a tool for epidemiologists to unravel complex associations previously intractable to robust nutritional research.

Materials and Methods

Study Populations. Samples for the primary study were drawn from three clinical trials and one sample biorepository, all based at Duke University in Durham, NC. The clinical trials consisted of a behavioral intervention that returned gut microbiome data to participants (NCT04037306, here "Weight Loss") and studies assessing the impact of fiber supplementation on the gut microbiota (NCT03595306, "Adult-1") (37) and on human cognition, behavior, and physiology (NCT04055246, "Adult-2") (38). The sample biorepository was collected from adolescents with obesity and their healthy-weight siblings (NCT02959034, "Adolescent") (62). Samples from an additional clinical trial that assessed the effects of baking soda supplementation in the context of a controlled diet (NCT02427594, "Controlled Feeding") (23) served as validation for the per-taxon accuracy analysis in the Weight Loss cohort. Application of *trnL* metabarcoding was a secondary analysis and determined exempt by the Duke Health Institutional Review Board (Pro00100567). Study characteristics and participant demographics are summarized in Table 1. All participants in each trial provided written informed consent and authorized future use of their deidentified stool samples for research.

Stool Sample Collection, Processing, and DNA Extraction. For all samples, *trnL* metabarcoding was performed on extracted fecal DNA originally generated for 16S rRNA microbiome sequencing. Stool samples were collected, stored, and DNA extracted as part of each primary study protocol. In all studies, stool was immediately frozen on collection by participants and transported frozen to a laboratory freezer. DNA extraction relied on versions of the PowerSoil kit system (Qiagen) following the manufacturer's instructions. Briefly, Weight Loss, Adult-1, and Adult-2 samples were extracted with the DNeasy PowerSoil or MagAttract PowerSoil kit, depending on number of samples per processing batch; Adolescent samples were extracted with the DNeasy PowerSoil Pro kit; and Controlled Feeding samples were extracted with the MagAttract PowerSoil kit. For Weight Loss, Adult-1, and Adult-2 samples, 1 to 1.5 g of stool was slurried in phosphate-buffered saline at a 10% weight-to-volume ratio in sterilized filter bags with a 0.33-mm pore size (Whirl-Pak) using a Stomacher 80 Biomaster (Seward Limited). Then, 750 μL of the slurry was added to tube-based (PowerSoil) and 200 μL to plate-based (PowerSoil MagAttract) extractions. For Adolescent samples, whole stool was added directly to the extraction; for Controlled Feeding samples, the tip cut from a stool swab was added directly to the extraction. Extracted DNA within each cohort was randomized and stored at −20 °C prior to *trnL* metabarcoding.

***trnL* Metabarcoding.** We performed *trnL* metabarcoding using a two-step PCR protocol. Primary PCR amplification of *trnL* used the KAPA HiFi HotStart PCR kit (KAPA Biosystems) in a 10-μL volume containing 0.5 μL of 10 μM forward and reverse primers (IDT), 2 μL of 5X KAPA HiFi buffer, 0.3 μL of 10 mM dNTPs, 0.1 μL of 100X SYBR Green I (Life Technologies), 0.1 μL KAPA HiFi polymerase, 3.5 μL nuclease-free water, and 3 μL of extracted DNA template. The primers were *trnL*(UAA) *g* and *h* (11) with Illumina overhang adapter sequences added at the 5′ end (*SI Appendix*, Table S1). Cycling conditions were an initial denaturation at 95 °C for 3 min, followed by 35 cycles of 98 °C for 20 s, 63 °C for 15 s, and 72 °C for 15 s. Each PCR batch included a positive and negative control, and samples were only advanced to the secondary PCR if controls performed as expected (otherwise, the entire batch was repeated). Secondary PCR amplification to add Illumina adapters and dual 8-bp indices for sample multiplexing was performed in a 50-μL volume containing 5 μL of 2.5 μM forward and reverse indexing primers (*SI Appendix*, Table S1), 10 μL of 5X KAPA HiFi buffer, 1.5 μL of 10 mM dNTPs, 0.5 μL of 100X SYBR Green I, 0.5 μL KAPA HiFi polymerase, 22.5 μL nuclease-free water, and 5 μL of primary PCR product diluted 1:100 in nuclease-free water.

Sequencing Library Preparation. Amplicons were cleaned (Ampure XP, Beckman Coulter), quantified (QuantIT dsDNA assay kit, Invitrogen), and combined in equimolar ratios to create a sequencing pool. If samples could not contribute enough DNA to fully balance the pool due to low post-PCR DNA concentration, they were added up to a set volume, typically 15 to 20 μL. Libraries were then concentrated, gel purified, quantified by both fluorimeter and qPCR, and spiked with 30% PhiX (Illumina) to mitigate low nucleotide diversity. Paired-end sequencing was carried out on an Illumina MiniSeq system according to the manufacturer's instructions using a 300-cycle Mid, 300-cycle High, or 150-cycle High kit (Illumina), depending on the number of samples in each pool. Due to the short length of *trnL* (median 89 bp in the reference, range 59 to 154 bp), 300-cycle kits guaranteed overlap between forward and reverse reads; 150-cycle kits did so for all but one plant, the water chestnut (*Eleocharis dulcis*).

Reference Database Construction. A list of edible plant taxa was compiled from US food availability data (63), global surveys (33), and reference volumes (64). DNA sequences likely to contain *trnL* were downloaded from two sources within the National Center for Biotechnology Information (NCBI): GenBank (all publicly available DNA sequence submissions) and the organelle genome resources of RefSeq (a curated, nonredundant subset of assembled chloroplast genomes). To obtain GenBank sequences, we used the entrez_search function of rentrez v1.2.3 (65) to submit separate queries for sequences containing "*trnL*" in any metadata field and each plant taxon name in the Organism field (*e.g.*, "*Zea mays*[ORGN] AND *trnL*" to pull data for corn, or *Z. mays*). Sequences with an "UNVERIFIED:" flag were discarded. To obtain RefSeq sequences, the plastid sequence release current as of June 2021 was downloaded and subset to only those accessions including an edible taxon name. Results from either source were then filtered to sequences containing primer binding sites for *trnL*(UAA)*g* and *trnL*(UAA)*h* in the correct orientation. The locations of primer binding sites within the parent sequence were identified using a custom R script with a mismatch tolerance of 20% [≤3 mismatches for *trnL*(UAA)*g* and ≤4 for *trnL*(UAA)*h*], and sequence outside the primer binding sites removed. Identical *trnL* sequences from different accessions of the same taxon were deduplicated, but we preserved distinct *trnL* sequences within taxa (indicating genetic variability) and identical *trnL* sequences from different taxa (indicating genetic conservation) to yield the final reference. The taxonomic tree of possible identifications in comparison to

the plant food phylogeny (Fig. 1*B*) was visualized with ggtree v. 2.2.4 (66), and taxa were mapped to "major" or "minor" crop labels following (67).

Bioinformatic Analysis. For each sequencing run, raw sequencing data were demultiplexed using bcl2fastq v2.20.0.422 (68). Read-through into the Illumina adapter sequence at the 3′ end was detected and right-trimmed with BBDuk v. 38.38 (69). Using cutadapt v. 3.4 (70), paired reads were filtered to those beginning with the expected primer sequence [either *trnL*(UAA)*g* for the forward read or *trnL*(UAA)*h* for the reverse] and then trimmed of both 5′ and 3′ sequences using a linked adapter format with a 15% error tolerance. Paired reads were quality-filtered by discarding reads with >2 expected errors and truncated at the first base with a quality score ≤ 2, denoised, and merged to produced ASVs using DADA2 v. 1.10.0 (71). We subsequently converted the pipeline to use existing infrastructure for amplicon marker gene analysis maintained in QIIME2 (72), with paired-end adapter and primer trimming performed with cutadapt v. 4.1 and sequences denoised with DADA2 v. 1.22.0. For future users, we recommend the QIIME2 pipeline for its simplicity, built-in tools for reproducibility and data provenance, and support of modular plugins that can facilitate further development of the ways in which *trnL* data are analyzed.

Taxonomic assignment was done with DADA2's assignSpecies function, which identified ASVs by exact sequence matching to the custom *trnL* reference database, with multiple matches allowed. If multiple matches occurred, reads were assigned to the taxon representing the last common ancestor of all matched taxa [*e.g.*, an ASV matching to both wheat (*Triticum aestivum*) and rye (*Secale cereale*) was relabeled as Poaceae, the family shared by both genera]. Sequence data were screened for contamination on a per-PCR batch basis using decontam v1.8.0 (73) using DNA quantitation data from the library pooling step, and suspected contaminants were removed. ASV count tables, taxonomic assignments, and metadata were organized using phyloseq v1.32.0 (74).

Prior to calculating pMR, ASVs identified to the same food taxon were automatically merged to make pMR representative of food identity, rather than *trnL* sequence variation. We reasoned that this was a natural mapping to existing metrics of dietary diversity, which measure foods recognized as distinct by the consumer. A small number of ASVs ($n = 4$) representing distinct sets of species within the same taxonomic label, which occur due to the last common ancestor method above, were identified and preserved in the data (*e.g.*, in the family Rosaceae, the rosids, one sequence variant indicates apple and pear intake and a second identifies strawberries and raspberries, so these were not merged). pMR was then calculated as the number of unique taxa observed with at least one read count in each sample. Shannon diversity was calculated using the diversity function from vegan v2.6 (75). Faith's PD was calculated using the function pd from picante v1.8.2 (76) on a tree constructed by aligning detected *trnL*-P6 ASVs, making a neighbor-joining (NJ) tree, and then fitting a generalized time-reversible with Gamma rate variation maximum likelihood tree using the NJ tree as a starting point. The tree was rooted using spirulina (*Arthrospira platensis*), a cyanobacterial species that is detectable with *trnL*, as an outgroup. Ideally, PD would be computed based on phylogenetic relationships inferred from whole-genome sequences; however, this was not possible due to the absence of some detected food plants from even an extensive curated food plant phylogeny [Fig. 1*B*; $n = 19$ (13%) of detected foods not included in the tree].

Dietary Data Collection and Processing.

Digital menus (Weight Loss). Complete menu data for each participant were exported from RealChoices menu software (SciMed Solutions) and linked to ingredient names from recipe source files. Ingredient common names were then manually identified to plant species using the NCBI Taxonomy Browser and Integrated Taxonomic Information System databases. For ingredients that were themselves composite foods (*e.g.*, "whole wheat bread"), we identified a primary ingredient using either provided brand information or the USDA FoodData Central database, which includes taxon mapping under the "Other information" header.

For all foods, portion sizes were estimated with FoodData Central by converting the recorded menu amount (*e.g.*, teaspoon, cup, ounce-weight, and slice) to a gram amount using the average weights under the "Measures" header.

Interventional diet records (Controlled Feeding). Diets were adapted from the control arm of prior feeding trials and designed using Food Processor software to be similar to typical US intake. Four daily menus were selected and scaled to calorie needs of participants to maintain stable body weight. Menus recurred in exact order during week 1 and week 2 of the feeding period to create directly comparable dietary exposure at both timepoints when stool was ascertained. All ingredients were purchased by the study team and prepared at the metabolic kitchen at the Stedman Center for Nutrition and Metabolism. Participants were fed their largest meal of the day weekdays on site under direct supervision by the study team with additional foods and beverages packed out for home consumption. Uneaten foods were returned along with daily dietary logs including additionally consumed food and beverage items. Dietary intake was assessed in Food Processor including any additionally consumed foods. Complete dietary data were provided in records kept by the primary study team, which accounted for differences between provisioned foods and returned, uneaten items. Dietary coding was performed as for the Weight Loss digital menus and by the same individual for consistency.

Dietary surveys (Adult-1 and Adult-2). Habitual dietary intake over the past 1 mo was assessed by administration of National Cancer Institute Diet History Questionnaire III (DHQ3), a 135-item, semiquantitative FFQ. FFQ data were quality checked by estimating participant basal metabolic rate (BMR) using the Harris–Benedict equation (77), calculating the ratio of reported calorie intake to estimated BMR, and excluding FFQs where this ratio was ≥2 absolute deviations outside the median of the full dataset (corresponding to a ratio of <0.22 or >1.75) from further analysis, as done in a prior study (44). This criterion preserved 87% of completed FFQs in the dataset (excluded responses were all for suspected underreporting).

Food Variety Score (FVS). The FVS was calculated as the number of unique food items consumed at least once per week. After summing daily intake frequencies within each food item, we tallied items with a daily frequency of consumption ≥0.14 [equivalent to 1/7, or a weekly frequency, as previously done for calculating FVS from frequency data (39)]. The plant component of the overall FVS was calculated using the same procedure after manually labeling food items derived from plants or including a plant ingredient. Total and plant component FVS were then adjusted for overall calorie intakes using the nutrient residual method (78): briefly, a linear regression model was used to fit FVS to overall energy intake in kilocalories, and the residuals from the model were used in place of raw FVS values.

Healthy Eating Index 2015 (HEI-2015). The HEI-2015 and its component scores were calculated automatically by the DHQ3. We defined a plant HEI score as the sum of exclusively plant-based adequacy components (Total Vegetables, Greens and Beans, Total Fruits, Whole Fruits, and Whole Grains), which give higher scores to higher intakes of encouraged plant food groups. Conversely, we defined a non-plant-based HEI score as the sum of components with exclusively non-plant-based items (Dairy, Sodium, Added Sugars, and Saturated Fat). Saturated Fat may contain plant items like palm oil or coconut, but we expect this category is largely reflective of meat and dairy intake. Though meat and seafood are included in HEI component scores, their categories also include plant-based items (legumes for "Total Protein" and legumes, nuts, seeds, and soy for "Seafood and Plant Protein"). We therefore did not include these categories in either score definition above.

Plant-based dietary index (PDI). The PDI and its variations, healthy PDI (hPDI) and unhealthy PDI (uPDI), were calculated from DHQ3 data by manually assigning food items to specified food groups ($n = 18$), splitting participants into quintiles based on gram weight of intake of each food group, and then scoring the quintiles from either 5 to 1 or 1 to 5, depending on the index being calculated. Food group scores were then summed within each participant to give the overall score. In rare cases, enough participants did not report consuming the food that they could not all be accommodated by the first quintile of the data; in this case, all participants with zero intake were assigned to the first quintile, and the remainder of the data split into quartiles and assigned to the second to fifth intake categories.

Statistical Analysis.

Interventional cohorts. For the per-taxon accuracy analysis in the Weight Loss cohort, we compared *trnL* presence or absence to the presence or absence of the same food taxon in the menu record from 1 to 2 d prior. The 1 to 2 d window was selected to account for the mean (28 h) and typical variation of measured gastrointestinal transit times in humans (34, 35). Responses were coded as true positives (TP, food present by both *trnL* and menu), true negatives (TN, absent by both *trnL* and menu), false positives (FP, present by *trnL*, not by menu), and false negatives (FN, absent by *trnL*, but present in menu). Accuracy was calculated as the percentage of true detections out of all detections or (TP + TN)/(TP + TN + FP + FN). We performed the same analysis in the Controlled Feeding cohort and calculated false-positive and false-negative error rates for the subset

of taxa ($n = 42$) that occurred in both Weight Loss and Controlled Feeding. The false-positive rate (FPR) was calculated as the percentage of samples with no record of prior intake where the food item was nevertheless detected by *trnL* or FP/(FP + TN); the false-negative rate (FNR) was calculated as the percentage of samples with recorded intake events that did not have the food detected by *trnL* or FN/(FN + TP). Two-tailed Spearman correlations between FPR and FNR in the two studies were performed using the cor.test function from R stats v4.1.3 (as is the case for all subsequent correlations in our statistical analysis).

For the portion size analysis in the Weight Loss cohort, we compared the centered log-ratio (CLR) transform of *trnL* read count to the quantity of the same food taxon estimated in the menu record from 1 to 2 d prior. We considered either continuous (weight in grams) or categorical (tertiles of serving size) measures of portion size. Serving size tertiles were estimated by calculating recorded serving sizes of each food taxon over 2 d spans (to match the estimated *trnL* transit time window) for all participants in the menu record and then assigning them to a tertile using the quantile function in R stats v4.1.3.

Two-tailed Spearman correlations were used to test for association between pMR and counts of plant taxa from menu records. The number of unique plants recorded in the menu was averaged from the 2 d preceding each stool sample. As a negative control, we also paired pMR with the two menu days following sample collection or two consecutive menu days chosen at random from the full dataset. Menu data from Saturdays and Sundays were excluded from the paired analysis because the on-site cafeteria only provided breakfast on weekends, and digital menus had to be supplemented with less accurate self-reports. As a result, a small number of samples were excluded from the primary test ($n = 4$, collected on Mondays) or included with only one comparison menu day ($n = 7$ collected on Sundays or Tuesdays). Outliers were identified by calculating the median difference between the number of plant taxa recorded in the menu from the number detected by *trnL* metabarcoding, and labeling outliers as samples with a difference ≥2 median absolute deviations outside that of the full dataset. Serving size and diversity analyses were not replicated in the Controlled Feeding cohort because the menu, by design, did not include large portions or variety of plant foods.

Adult-1 and Adult-2 cohorts. Two-tailed Spearman correlations were calculated between mean pMR (averaged across all samples for each participant) and FFQ data. For each subsampling scheme, samples that fit each strategy were randomly selected from the total available for each participant, and Spearman correlations were calculated using the mean pMR of only those samples. One hundred subsampling iterations were performed for each scheme, unless fewer unique combinations were available or duplicate subsamples occurred by chance (this resulted in a loss of no more than three iterations from any combination of study, dietary index, and sampling scheme).

Adolescent cohort. Demographic, health, and socioeconomic status variables were included as covariates in a linear model with pMR as the outcome variable. All covariates were checked for completeness and missing entries coded as "Unknown" ($n = 62$ for income and $n = 28$ for food insecurity) so as not to exclude missing data. We chose not to impute missing values because we hypothesized that missing responses to socioeconomic questions likely violated assumptions that data are missing completely at random (*i.e.*, individuals in lower income or food-insecure categories would be more likely to leave the question blank). "Unknown" categories were included in the model but not visualized in Fig. 3*B* (their fitted coefficients are reported in *SI Appendix*, Table S2). Because only 138 of 246 subjects (56%) had two timepoints, we used a linear model of pMR from the "Entry" timepoint alone rather than a mixed-effects model with repeated measurements. The distribution of pMR was approximately normal (tested with the descdist function of fitdistrplus v1.1.8), so we tested both a linear model using the lm function of R stats v. 4.1.3 and a negative binomial family generalized linear model (GLM) using the glm function, which as a discrete distribution is a theoretically better approximation of pMR. Both yielded similar results, and we present the findings of the linear model here for simpler interpretation of the magnitude of fitted coefficients. We screened for, but did not detect, collinearity among model predictors using the function vif of car package v3.0.12. Observed versus predicted pMR and residual versus predicted pMR plots were generated to check model validity.

Rarefaction. Rarefaction was performed using vegan v2.5.7, and statistical tests above were repeated using rarefied pMR in place of raw pMR. Rarefaction provides a statistical estimate of richness that adjusts for variation in sequencing depth, which we first noted in the Adult-1 and Adult-2 cohorts (range 1 to 150,330, *SI Appendix*, Fig. S10*A*) despite experimental strategies to balance samples within each sequencing batch. Because richness scales with sampling effort (79) (*SI Appendix*, Fig. S10*B*), we tested whether using rarefaction (statistical downsampling to a shared read depth) to adjust for differences in sequencing depth affected relationships between pMR and dietary data. Rarefaction strengthened the correlation between pMR and recorded menus in the Weight Loss cohort; in the Adult-1 and Adult-2 cohorts, rarefaction retained the positive correlations to FVS and hPDI at only slightly weakened magnitude but rendered the relationship to HEI-2015 plant component score insignificant (*SI Appendix*, Table S2). One interpretation of these findings is that read depth may indicate plant content of the diet rather than technical variation in sample preparation. In support of this hypothesis, FVS plant residuals, overall PDI, and HEI-plant component scores were all significantly lower for Adult-1 and Adult-2 samples with fewer than 1,000 reads, indicating reduced plant intake by an independent measure (*SI Appendix*, Fig. S10*C*). Therefore, we continued subsequent analyses without rarefaction (while monitoring its effects in *SI Appendix*, Table S2).

Data, Materials, and Software Availability. Raw, demultiplexed *trnL* sequencing data are deposited to the European Nucleotide Archive under the accessions PRJEB62684 (Weight Loss) (80), PRJEB62685 (Adult-1) (81), PRJEB62686 (Adult-2) (82), and PRJEB62687 (Adolescent) (83). Deidentified clinical metadata associated with this study are available upon request and will be shared when consistent with applicable study agreements, regulations, and ethical standards. Code associated with this manuscript is organized into two repositories available on Zenodo: (1) the bioinformatic pipeline and reference database to analyze raw trnL sequencing data, accompanied by a test dataset and tutorial (https://zenodo.org/record/8004348) (84) and (2) R scripts to reproduce manuscript results from processed trnL sequencing data (https://zenodo.org/record/8004413). from GitHub (85).

ACKNOWLEDGMENTS. We thank our study volunteers for their participation; Verónica Palacios for human study support; Michelle Kirtley for manuscript edits; Tyler Kartzinel for experimental and conceptual insights; and Tonya Snipes, Lisa Alston-Latta, and Margaret Huggins for keeping our lab spaces and glassware clean. Funding for this work came from the National Institute of Diabetes and Digestive and Kidney Diseases (grants R24-DK110492, R01-DK116187, and R01-DK128611), the Burroughs Wellcome Fund Pathogenesis of Infectious Disease Award, the Duke Microbiome Center, the Springer Nature Limited Global Grant for Gut Health, the Chan Zuckerberg Initiative, the Triangle Center for Evolutionary Medicine, the Integrative Bioinformatics for Investigating and Engineering Microbiomes Graduate Student Fellowship, and the Ruth L. Kirschstein National Research Service Award to the Duke Medical Scientist Training Program. This work used a high-performance computing facility partially supported by grants 2016-IDG-1013 ("HARDAC+: Reproducible HPC for Next-generation Genomics") and 2020-IIG-2109 ("HARDAC-M: Enabling memory-intensive computation for genomics") from the North Carolina Biotechnology Center.

Author affiliations: [a]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC 27710; [b]Medical Scientist Training Program, Duke University School of Medicine, Durham, NC 27710; [c]Duke Microbiome Core Facility, Center for Genomic and Computational Biology, Duke University, Durham, NC 27710; [d]Duke Lifestyle and Weight Management Center, Durham, NC 27710; [e]Department of Medicine, Duke University School of Medicine, Durham, NC 27710; [f]Department of Medicine, Nephrology Division, Sarah W. Stedman Nutrition and Metabolism Center, Duke University Medical Center, Durham, NC 27705; [g]Department of Medicine, University of Virginia School of Medicine, Charlottesville, VA 22903; [h]Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA 22903; [i]Division of Pediatric Infectious Diseases, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL 60611; [j]Duke Microbiome Center, Duke University School of Medicine, Durham, NC 27710; [k]Department of Pediatrics, Duke University School of Medicine, Durham, NC 27710; and [l]Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

1. V. Kipnis *et al.*, Structure of dietary measurement error: Results of the OPEN Biomarker Study. *Am. J. Epidemiol.* **158**, 14–21 (2003).
2. A. F. Subar *et al.*, Addressing current criticism regarding the value of self-report dietary data. *J. Nutr.* **145**, 2639–2645 (2015).
3. F. E. Thompson *et al.*, The National Cancer Institute's Dietary Assessment Primer: A resource for diet research. *J. Acad. Nutr. Diet.* **115**, 1986–1995 (2015).
4. W. Willett, *Nutritional Epidemiology* (Oxford University Press, 2012).
5. A. B. Ross *et al.*, Plasma alkylresorcinols as a biomarker of whole-grain food consumption in a large population: Results from the WHOLEheart Intervention Study. *Am. J. Clin. Nutr.* **95**, 204–211 (2012).
6. M. Guasch-Ferré, S. N. Bhupathiraju, F. B. Hu, Use of metabolomics in improving assessment of dietary intake. *Clin. Chem.* **64**, 82–98 (2018).

7. M. C. Playdon *et al.*, Identifying biomarkers of dietary patterns by using metabolomics. *Am. J. Clin. Nutr.* **105**, 450–465 (2017).
8. Diet History Questionnaire III (DHQ III). https://epi.grants.cancer.gov/dhq3/index.html. Accessed 13 January 2023.
9. F. E. Thompson, A. F. Subar, "Dietary assessment methodology" in *Nutrition in the Prevention and Treatment of Disease*, A. M. Coulston, C. Boushey, M. G. Ferruzzi, L. M. Delahanty, Eds. (Academic Press, ed. 4, 2017), pp. 5–48.
10. T. R. Kartzinel *et al.*, DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8019–8024 (2015).
11. P. Taberlet *et al.*, Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* **35**, e14–e14 (2007).
12. A. T. Reese *et al.*, Using DNA metabarcoding to evaluate the plant component of human diets: A proof of concept. *mSystems* **4**, e00458–19 (2019).
13. J. Schneider *et al.*, Comprehensive coverage of human last meal components revealed by a forensic DNA metabarcoding approach. *Sci. Rep.* **11**, 8876 (2021).
14. Y. Choi, S. L. Hoops, C. J. Thoma, A. J. Johnson, A guide to dietary pattern–microbiome data integration. *J. Nutr.* **152**, 1187–1199 (2022).
15. J. Zhao *et al.*, A review of statistical methods for dietary pattern analysis. *Nutr. J.* **20**, 37 (2021).
16. H. C. Boshuizen, D. E. te Beest, Pitfalls in the statistical analysis of microbiome amplicon sequencing data. *Mol. Ecol. Resour.* **23**, 539–548 (2022).
17. R. H. Walker *et al.*, Mechanisms of individual variation in large herbivore diets: Roles of spatial heterogeneity and state-dependent foraging. *Ecology* **104**, e3921 (2023).
18. A. C. Thomas, B. E. Deagle, J. P. Eveson, C. H. Harsch, A. W. Trites, Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Mol. Ecol. Resour.* **16**, 714–726 (2016).
19. D. M. Gohl *et al.*, Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **34**, 942–949 (2016).
20. A. F. Subar *et al.*, Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires. *Am. J. Epidemiol.* **154**, 1089–1099 (2001).
21. R. Milla, Crop Origins and Phylo Food: A database and a phylogenetic tree to stimulate comparative analyses on the origins of food crops. *Global Ecol. Biogeogr.* **29**, 606–614 (2020).
22. B. J. Callahan, P. J. McMurdie, S. P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).
23. C. C. Tyson *et al.*, Effect of bicarbonate on net acid excretion, blood pressure, and metabolism in patients with and without CKD: The Acid Base Compensation in CKD study. *Am. J. Kidney Dis.* **78**, 38–47 (2021).
24. USDA/HHS, "Nutrition and your health: Dietary Guidelines for Americans" (1980).
25. A. Hatløy, J. Hallund, M. M. Diarra, A. Oshaug, Food variety, socioeconomic status and nutritional status in urban and rural areas in Koutiala (Mali). *Public Health Nutr.* **3**, 57–65 (2000).
26. M. T. Ruel, Operationalizing Dietary Diversity: A Review of Measurement Issues and Research Priorities. *J. Nutr.* **133**, S3911–S3926 (2003).
27. N. P. Steyn, J. Nel, D. Labadarios, E. M. W. Maunder, H. S. Kruger, Which dietary diversity indicator is best to assess micronutrient adequacy in children 1 to 9 y? *Nutrition* **30**, 55–60 (2014).
28. A. K. Kant, G. Block, A. Schatzkin, R. G. Ziegler, M. Nestle, Dietary diversity in the US population, NHANES II, 1976–1980. *J. Am. Diet Assoc.* **91**, 1526–1531 (1991).
29. G. Kennedy, T. Ballard, M. C. Dop, *Guidelines for Measuring Household and Individual Dietary Diversity* (Food and Agriculture Organization of the United Nations, 2011).
30. F. B. Hu, Dietary pattern analysis: A new direction in nutritional epidemiology. *Curr. Opin. Lipidol.* **13**, 3–9 (2002).
31. M. C. de Oliveira Otto *et al.*, Dietary diversity: Implications for obesity prevention in adult populations: A science advisory from the American Heart Association. *Circulation* **138**, e160–e168 (2018).
32. E. O. Verger *et al.*, Dietary diversity indicators and their associations with dietary adequacy and health outcomes: A systematic scoping review. *Adv. Nutr.* **12**, 1659–1672 (2021).
33. C. Lachat *et al.*, Dietary species richness as a measure of food biodiversity and nutritional quality of diets. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 127–132 (2018).
34. G. K. Nandhra *et al.*, Normative values for region-specific colonic and gastrointestinal transit times in 111 healthy volunteers using the 3D-Transit electromagnet tracking system: Influence of age, gender, and body mass index. *Neurogastroenterol. Motil.* **32**, e13734 (2020).
35. F. Asnicar *et al.*, Blue poo: Impact of gut transit time on the gut microbiome using a novel marker. *Gut* **70**, 1665–1674 (2021).
36. Centers for Disease Control and Prevention, National Center for Health Statistics, NHANES - What We Eat in America. *What We Eat in America, DHHS-USDA Dietary Survey Integration* (2020). https://www.cdc.gov/nchs/nhanes/wweia.htm. Accessed 6 March 2023.
37. Z. C. Holmes *et al.*, Microbiota responses to different prebiotics are conserved within individuals and associated with habitual fiber intake. *Microbiome* **10**, 114 (2022).
38. J. Letourneau *et al.*, Ecological memory of prior nutrient exposure in the human gut microbiome. *ISME J.* **16**, 2479–2490 (2022).
39. G. Masset, P. Scarborough, M. Rayner, G. Mishra, E. J. Brunner, Can nutrient profiling help to identify foods which diet variety should be encouraged? Results from the Whitehall II cohort. *Br. J. Nutr.* **113**, 1800–1809 (2015).
40. E. A. Hu, L. M. Steffen, J. Coresh, L. J. Appel, C. M. Rebholz, Adherence to the Healthy Eating Index–2015 and other dietary patterns may reduce risk of cardiovascular disease, cardiovascular mortality, and all-cause mortality. *J. Nutr.* **150**, 312–321 (2020).
41. C. Panizza *et al.*, Testing the predictive validity of the Healthy Eating Index-2015 in the Multiethnic Cohort: Is the score associated with a reduced risk of all-cause and cause-specific mortality? *Nutrients* **10**, 452 (2018).
42. A. Satija *et al.*, Plant-based dietary patterns and incidence of type 2 diabetes in US men and women: Results from three prospective cohort studies. *PLoS Med.* **13**, e1002039 (2016).
43. A. Satija *et al.*, Healthful and unhealthful plant-based diets and the risk of coronary heart disease in U.S. *adults. J. Am. College of Cardiol.* **70**, 411–422 (2017).
44. F. Asnicar *et al.*, Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).
45. A. Drewnowski, S. A. Renderson, A. Driscoll, B. J. Rolls, The Dietary Variety Score: Assessing diet quality in healthy young and older adults. *J. Am. Diet. Assoc.* **97**, 266–271 (1997).
46. M. W. Gillman, Family dinner and diet quality among older children and adolescents. *Arch. Family Med.* **9**, 235–240 (2000).
47. A. Salehi-Abargouei, F. Akbari, N. Bellissimo, L. Azadbakht, Dietary diversity score and obesity: A systematic review and meta-analysis of observational studies. *Eur. J. Clin. Nutr.* **70**, 1–9 (2016).
48. M. Vadiveloo, L. B. Dixon, N. Parekh, Associations between dietary variety and measures of body adiposity: A systematic review of epidemiological studies. *Br. J. Nutr.* **109**, 1557–1572 (2013).
49. M. C. de Oliveira Otto, N. S. Padhye, A. G. Bertoni, D. R. Jacobs, D. Mozaffarian, Everything in moderation - Dietary diversity and quality, central obesity and risk of diabetes. *PLoS One* **10**, e0141341 (2015).
50. M. A. McCrory *et al.*, Dietary variety within food groups: Association with energy intake and body fatness in men and women. *Am. J. Clin. Nutr.* **69**, 440–447 (1999).
51. M. Fanelli Kuczmarski *et al.*, Aspects of dietary diversity differ in their association with atherosclerotic cardiovascular risk in a racially diverse US adult population. *Nutrients* **11**, 1034 (2019).
52. J. Hoddinott, Y. Yohannes "Dietary diversity as a food security indicator" (International Food Policy Research Institute (IFPRI), 2002).
53. L. K. Khan, R. Martorell, Diet diversity in Mexican Americans, Cuban Americans and Puerto Ricans. *Ecol. Food Nutr.* **36**, 401–415 (1997).
54. A. Kant, A. Schatzkin, T. Harris, R. Ziegler, G. Block, Dietary diversity and subsequent mortality in the first national health and nutrition examination survey epidemiologic follow-up study. *Am. J. Clin. Nutr.* **57**, 434–440 (1993).
55. L. E. Torheim *et al.*, Nutrient adequacy and dietary diversity in rural Mali: Association and determinants. *Eur. J. Clin. Nutr.* **58**, 594–604 (2004).
56. I. N. Bezerra, R. Sichieri, Household food diversity and nutritional status among adults in Brazil. *Int. J. Behav. Nutr. Phys. Act.* **8**, 22 (2011).
57. M. Nesbitt, R. P. H. McBurney, M. Broin, H. J. Beentje, Linking biodiversity, food and nutrition: The importance of plant identification and nomenclature. *J. Food Compos. Anal.* **23**, 486–498 (2010).
58. C. L. Schoch *et al.*, Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi*. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6241–6246 (2012).
59. K. Hadziavdic *et al.*, Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS ONE* **9**, e87624 (2014).
60. Ş Procheş, J. R. U. Wilson, J. C. Vamosi, D. M. Richardson, Plant diversity in the human diet: Weak phylogenetic signal indicates breadth. *BioScience* **58**, 151–159 (2008).
61. Q. D. Read, K. L. Hondula, M. K. Muth, Biodiversity effects of food system sustainability actions from farm to fork. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2113884119 (2022).
62. J. R. McCann *et al.*, The pediatric obesity microbiome and metabolism study (POMMS): Methods, baseline data, and early insights. *Obesity* **29**, 569–578 (2021).
63. Economic Research Service (ERS) of the U.S. Department of Agriculture (USDA), USDA ERS - Food Availability (Per Capita) Data System. https://www.ers.usda.gov/data-products/food-availability-per-capita-data-system/. Accessed March 9, 2023.
64. B.-E. Van Wyk, *Food Plants of the World: An Illustrated Guide* (Timber Press, ed. 1, 2005).
65. J. D. Winter, rentrez: An R package for the NCBI eUtils API. *The R Journal* **9**, 520 (2017).
66. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T. Lam, ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
67. R. Milla, C. P. Osborne, Crop origins explain variation in global agricultural relevance. *Nat. Plants* **7**, 598–607 (2021).
68. Illumina. bcl2fastq Conversion Software. https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html. Accessed 1 August 2022.
69. B. Bushnell. BBTools software package. http://bbtools.jgi.doe.gov. Accessed 18 April 2018.
70. M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10 (2011).
71. B. J. Callahan *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583 (2016).
72. E. Bolyen *et al.*, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
73. N. M. Davis *et al.*, Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
74. P. J. McMurdie, S. Holmes. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
75. J. Oksanen *et al.*, vegan: *Community Ecology* Package. R package version 2.6, https://CRAN.R-project.org/package=vegan (2022).
76. S. W. Kembel *et al.*, Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).
77. M. Mifflin *et al.*, A new predictive equation for resting energy expenditure in healthy individuals. *Am. J. Clin. Nutr.* **51**, 241–247 (1990).
78. W. Willett, G. Howe, L. Kushi, Adjustment for total energy intake in epidemiologic studies. *Am. J. Clin. Nutr.* **65**, S1220–S1228 (1997).
79. N. J. Gotelli, R. K. Colwell, Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391 (2001).
80. B. L. Petrone *et al.*, trnL sequencing data for Weight Loss cohort. *European Nucleotide Archive*. https://www.ebi.ac.uk/ena/browser/view/PRJEB62684. Accessed 30 May 2023.
81. B. L. Petrone, S. Jiang, H. K. Durand, L. A. David, trnL sequencing data for Adult-1 cohort. *European Nucleotide Archive*. https://www.ebi.ac.uk/ena/browser/view/PRJEB62685. Accessed 30 May 2023.
82. B. L. Petrone, S. Jiang, H. K. Durand, L. A. David, trnL sequencing data for Adult-2 cohort. *European Nucleotide Archive*. https://www.ebi.ac.uk/ena/browser/view/PRJEB62686. Accessed 30 May 2023.
83. B. L. Petrone *et al.*, trnL sequencing data for *Adolescent* cohort. *European Nucleotide Archive*. https://www.ebi.ac.uk/ena/browser/view/PRJEB62687. Accessed 30 May 2023.
84. B. L. Petrone, trnL metabarcoding protocols and bioinformatic pipeline (Version 1.0). *Zenodo*. https://zenodo.org/record/8004348. Accessed 5 June 2023.
85. B. L. Petrone, Data and code accompanying plant richness manuscript (Version 1.0). *Zenodo*. https://zenodo.org/record/8004413. Accessed 5 June 2023.